Electronic Scholarly Publishing: Foundations of Genetics

Robert J. Robbins

Fred Hutchinson Cancer Research Center

ABSTRACT

As the Human Genome Project (HGP) moves toward its successful completion, an increasing number of individuals are becoming interested in understanding this project and its results. Since the HGP has potentially significant ethical, legal, and social implications for all citizens, the number of individuals who do, or should wish to become familiar with the project is very high. In addition to its importance in the training of professional geneticists, the HGP is of special relevance for undergraduate training in basic biology, and even for high-school and other K-12 education.

Understanding the results of HGP research requires a familiarity with the notions of basic genetics. We propose to create an educational resource at which classic literature, establishing the foundations of modern genetics, will be republished in readily available, typeset-quality electronic form. Making the results of genome research widely available and accessible (both intellectually and physically) is a key goal of the ELSI component of the Human Genome Project. By providing interested parties worldwide with ready access to the intellectual foundations of genetics, the proposed electronic-publishing project will greatly facilitate the achievement of that goal.

Preliminary results from a demonstration project have established the feasibility of such a project and have offered some indications of great need and demand for such a project. In addition to publishing original literature, we will also provide a variety of tools to facilitate access and use of our publications. Our site will be designed to function in a federated information infrastructure and we will actively support the work of third-party, value-adding developers.

In 1900, genetics was just beginning as scientific discipline. As 2000 approaches, the Human Genome Project promises to deliver molecular-level information detailing our genetic heritage. Now, the early results of the HGP hold out great promise in areas from agriculture to biotechnology to medicine and beyond. As the HGP runs to completion, its effects will spread throughout society, stirring increasing interest among informed citizens. Understanding the significance of the HGP in part depends upon a basic understanding of genetic principles. This project will help bring that understanding to a wide audience.

Electronic Scholarly Publishing: Foundations of Genetics

Robert J. Robbins

Fred Hutchinson Cancer Research Center

INTRODUCTION

As the Human Genome Project (HGP) generates amazing results and garners positive publicity in the scientific, the commercial, and the lay press, an increasing number of individuals are becoming interested in understanding this project and its results. Since the HGP has potentially significant ethical, legal, and social implications for all citizens, the number of individuals who do, or should wish to become familiar with the project is very high. In addition to its importance in the training of professional geneticists, the HGP is of special relevance for undergraduate training in basic biology, and even for high-school and other K-12 education.

Understanding the results of HGP research requires a familiarity with the notions of basic genetics. Unlike other disciplines that evolved over centuries, modern genetics began abruptly with the rediscovery of Gregor Mendel's work, and within a few years, fundamental concepts were elaborated and the foundations of genetics established.

Because genetics developed so rapidly in just a few decades after 1900, the literature of that period constitutes a valuable resource even now. It may be read profitably by students and scientists wishing to understand the foundations of their field, as well as by laymen or historians of science. Unfortunately, the early literature is rapidly becoming almost inaccessible. Newer libraries do not hold older journals and even established libraries are moving their materials from that era into hard-to-reach (and impossible to browse) long-term storage in remote warehouses.

To be sure, key studies from the early work are discussed in nearly all textbooks, but a comparison of these presentations with the actual literature shows that most textbook treatments have essentially mythologized the early work so that real understanding is lost. There have been several collections of classic works developed over the years (although none lately), but these suffer from the effects of the necessary, but nonetheless pernicious, highly selective sampling that accompanies these projects. Such selectivity, coupled with introductions that offer essentially modern interpretations of the work, obscure the intellectual rigor and excitement of the original efforts.

To address these problems, we propose to republish classic literature in genetics in electronic form, so that access to these essential works will be trivially easy for all those with access to simple web-browsing software. A modest test version of such a publishing site has already been established as *Classic Papers in Genetics*:

`http://www.gdb.org/rjr/history.html`

Making the results of genome research widely available and accessible (both intellectually and physically) is a key goal of the ELSI component of the Human Genome

Project. By providing interested parties worldwide with ready access to the intellectual foundations of genetics, the proposed electronic-publishing project will greatly facilitate the achievement of that goal.

Proposal

We will establish an Electronic Scholarly Publishing (ESP) site that will republish approximately fifteen thousand pages of original literature in classical genetics in a readily accessible electronic form. The documents will be rendered as typeset-quality electronic files, in Adobe Acrobat PDF format and they will be made available via a World-Wide Web server maintained at the Fred Hutchinson Cancer Research Center (FHCRC).

The interface and accompanying tools will be designed to assist users in locating files of interest. Separate browse lists will be maintained for the entire collection, and for several subtopics within the collection (e.g., population genetics, physiological genetics, Drosophila genetics, etc.).Each browse list will be maintained in multiple copies, each in a different sort order: by author name, by title, by subject, by date of original publication, and by date of electronic republication (this "what's new" listing will help frequent visitors locate material added since their last visit.). In addition, a powerful search engine will provide full-text indexing on the content of the documents. Finally, a professional librarian/cataloger will catalog and classify all documents as they are republished. This cataloging will support both automated browse-list creation and user-driven database searching.

Many users now complain that locating relevant information on the WWW is becoming tedious and inefficient. The various tools deployed at the proposed site will alleviate many of these problems and should make locating relevant papers as easy as locating relevant books in a well-maintained research library.

Tools will also be provided to allow the selection and downloading of *sets* of files, so that instructors, for example, might retrieve blocks of papers for use in a local class. Tools will also be provided to allow the ready, almost automatic production of annotated copies of original documents so that instructors or scholars might easily prepare special annotated manuscripts either for local use, or for republication to the general scientific community.

A bibliographic database will maintain information about all republished documents, including cataloging and classification information. This database will be made generally available, so that third-party developers might add value by developing additional metatools, or so that bioinformatics experts at other sites might link the ESP database to their local systems.

Introductory comments will be provided for all republished documents. A board of editors will assist in selecting documents for inclusion, and may also contribute to drafting introductory materials or even scholarly essays.

Documents will be chosen for inclusion based on their scientific merit, which may not necessarily coincide with their currently being considered a "classic" paper. We will attempt to republish large blocks of related papers, including those that argued against positions that ultimately proved correct. Only by capturing the real contemporary dialogue can the excitement of true discovery be kept alive for readers.

The site will be useful not only to students, teachers, scholars, but also to general readers. The literature of classical genetics remains remarkably accessible to all readers, in part because all readers bring with them a ready familiarity with the *fact of inheritance* (like begets like, more or less) and need only to achieve an understanding of the *mechanisms of heredity* as a biological process.

Although modern molecular biology involves methods that can seem almost magical to the uninitiated, the original techniques of classical genetics are readily appreciated by one and all: cross individuals that differ in some inherited trait, collect all of the progeny, score their attributes, and propose mechanisms to explain the patterns of inheritance observed. Repeat as necessary.

In reading the early works of classical genetics, one is drawn, almost inexorably, into ever more complex models, until molecular explanations begin to seem both necessary and natural. At that point, the tools for understanding genome research are at hand. Assisting readers reach this point is the goal of this project.

Background

Although the problems of inheritance have fascinated observers since the beginning of recorded history, a systematic, scientific examination of heredity only began in the late nineteenth century, then exploded as an intellectual field shortly after 1900. The transformation of genetics from an observational, anecdotal endeavor to a true science with rigorous, deep theoretical foundations, and with profound explanatory power occurred very rapidly.

Classical Genetics, the study of patterns of inheritance, began with the rediscovery of Mendel's work in 1900 and flourished through the first half of the century. By 1950, however, classical genetics was a mature discipline and the driving unanswered question for genetics had become, *What is the chemical nature of the gene?* With the elucidation of DNA structure in 1953, molecular genetics began, culminating in the amazing successes of the Human Genome Project (HGP) nearly fifty years later.

Although the insights and results from modern molecular biology far outstrip the explanatory power of classical genetics, classical genetics still provides the biological context for understanding all of genetics. For this reason, a familiarity with classical genetics is important for biologists and lay persons alike, if they are to appreciate the findings of the HGP.

For several years, first at Johns Hopkins University and now at the Fred Hutchinson Cancer Research Center, I have been involved in teaching graduate courses that treat the history of the gene concept. These courses have involved extensive reading and discussion of early genetics literature. We examined topics in some detail, reading series of papers in full, not just isolated papers or selections from individual monographs. All of the students, including advanced graduate students working at the bench on state-of-the-art projects in molecular genetics, reacted enthusiastically to the course, both for its general intellectual stimulation and for the light that it shed upon modern thinking.

The immediate genesis of the proposed ESP project came from the experience in these graduate seminars, but the project's deeper roots trace back to my teaching experiences while at Michigan State University.

Pedagogical Issues

In the 1970s, a student once visited my office hours to ask about a lecture on the fundamental dogma of molecular biology (DNA makes RNA makes protein). Assuming that she wanted a more detailed explanation, I launched into an expanded presentation of the molecular facts. She listened attentively, but when I had finished, asked, "And you really believe all of that?" A recitation of scientific "facts" was not what she was after. Rather, she wanted to understand the process by which one comes to believe such things.

This encounter profoundly changed my approach to teaching biology, especially at an introductory level. Detailed presentations of molecular biological findings only show *what* we scientists believe. *Why* we believe remains inaccessible to students, unless we help them grasp the process of scientific investigation and reasoning. Textbooks and monographs offer excellent summaries of *what* we know, but really understanding *why* we believe requires contact with original literature.

The example of classical genetics provides powerful pedagogical tools for helping students understand the process of scientific investigation and the bass of scientific belief. Basic experiments in classical genetics can be appreciated by students with little or no formal scientific training. In these experiments, one crosses two individuals that differ in a single trait, counts the progeny, then draws inferences regarding the possible mechanisms of heredity. Although the first works are intellectually accessible to all, as the evidence mounts, the model becomes more complex, drawing one inexorably toward molecular explanations.

Although historical treatments run the risk of being seen as dry and dusty (especially if the "historical" treatment is merely a chronologically ordered recitation of facts), in my experience, if the foundations of classical genetics are presented with an eye on both the process and the excitement of discovery, students follow the material with great interest. Soon, they find themselves not only ready for, but demanding molecular explanations for genetic models they now are prepared both to understand and to believe.

For example, by 1950, classical genetic analysis had shown that the chemical gene, if there was such a thing, would have to possess two traits that seemed to be mutually exclusive:

1. The gene must be *heterocatalytic* – that is, it must be able to control the synthesis of other molecules of arbitrary complexity and detail.
2. The gene must be *autocatalytic* — that is, it must be able to control the synthesis of its own descendants with perfect fidelity.

Most surprisingly, the heterocatalylic function was known to be readily susceptible to mutation, yet the autocatalytic function was wholly resistant to mutation. Genes whose heterocatalytic function had been profoundly altered by mutation replicated as well as any normal gene.

These almost paradoxical requirements defined the necessary attributes of the chemical gene on the eve of the molecular revolution. By the early 1950's, evidence suggested that DNA might be the hereditary substance, but the current tetranucleotide model for the structure of DNA (a dull polymer of repeating identical tetramers) seemed to rule it out. Then, Watson and Crick proposed a model for DNA structure (Watson and Crick, 1953a) that transformed our thinking about biological molecules.

The significance of the Watson–Crick model for DNA structure can only be fully appreciated by someone familiar with the apparent paradox in the simultaneous requirement for hetero– and autocatalysis. Watson–Crick DNA is wholly unconstrained in one dimension (where four different nucleotides may be arrayed in any possible order along one strand), but totally constrained in another dimension (where base pairs in one strand must be perfectly complementary to their partners in the other strand).

Clearly, the heterocatalytic function must reside in the unconstrained linear ordering of nucleotides, while the autocatalytic function must reside in the totally constrained base-pairing between strands. Watson and Crick (1953b) explicitly noted this in their second paper:

> The phosphate-sugar backbone of our model is completely regular, but any sequence of the pairs of bases can fit into the structure. It follows that in a long molecule many different permutations are possible, and it therefore seems likely that the precise sequence of the bases is the code which carries the genetical information. If the actual order of the bases on one of the pair of chains were given, one could write down the exact order of the bases on the other one, because of the specific pairing. Thus one chain is, as it were, the complement of the other, and it is this feature which suggests how the deoxyribonucleic acid molecule might duplicate itself.

The point of this aside on the structure of DNA is to show that a grounding in classical genetics provides an excellent foundation for an appreciation of molecular genetics. And, the best way to gain a true scientific grounding in classical genetics is through reading the early literature.

Justification for the Proposal

The National Science Foundation has developed guidelines to assist applicants prepare proposals for its Database Activities in Biology program. The guidelines (which are applicable to all information-resource proposals), note that generally the most successful proposals allow reviewers to respond positively to each of the following questions: (1) Is there a real need for such an activity, (2) Will the proposed activity actually address that need, (3) Will the project likely be done, on time and within budget, and (4) Is it worth it? Here we address these issues.

Is there a need? Does it address the need?

The need for the proposed activity can be seen in the difficulty non-scientists have in comprehending molecular biology, combined with the increasing requirement that more and more non-scientists do just that in order to deal reasonably with the growing ethical, legal, and social implications of genome research, bioengineering, reproductive technologies, and other intrusions of biological knowledge into everyday life.

Documentation that the proposed activity will help in meeting that need comes from the world-wide response to a small demonstration project that has already been established to test the feasibility of the project. In 1995, a few papers were converted to PDF format and made available via a WWW page at Johns Hopkins. Even when only one or two papers were available, the site began to attract some attention and use. By February, 1995, the site had eight papers available and a test was begun to determine the level of use it would attract. Despite the fact that only eight papers were available, and despite the complete absence of any formal effort to publicize the site, during the thirteen-

week period ending 20 June, more than 700 file downloads were made by computers at 250 institutions in 30 different countries.

Although the majority of requests came from the United States, the diversity of requesting sites was striking, ranging from a Catholic girls school in New Zealand to the University of Tartu in Estonia. Intriguingly, in one session, all eight papers were downloaded to a machine whose domain (EOP.GOV) translated to "Executive Office of the President, Washington, DC, USA." More detailed information about all the sites requesting downloads from the demonstration project is given in Appendix I)

The eight papers and the number of requests for each are shown in Table I. Hardy's contribution to population genetics was the most popular, with Garrod's first paper on biochemical genetics a close second. Riddle's paper, certainly the least significant paper on the list, was requested least. The correlation of number of requests with intellectual significance of the papers suggests that users were motivated by more than mere curiosity.

Table I. Papers requested from the demonstration site over a thirteen week period spanning April - June, 1996.

Counts	Paper requested
142	Hardy, G. H. 1908. Mendelian proportions in a mixed population. *Science*, NS. XXVIII:49-50
120	Garrod, Archibald E. 1902. The incidence of alkaptonuria: A study in chemical individuality. *Lancet*, ii:1616-1620.
89	Morgan, Thomas, H. 1909. What are "factors" in Mendelian explanations? *American Breeders Association Reports*, 5:365-369.
88	Wright, Sewall. 1932. Complementary factors for eye color in *Drosophila*. *The American Naturalist*, LXVI:282-283.
87	Bridges, Calvin B. 1914. Direct proof through non-disjunction that the sex-linked genes of *Drosophila* are borne on the X-chromosome. *Science*, NS vol. XL:107-109.
84	Sutton, Walter S. 1902. On the morphology of the chromosome group in *Brachystola magna*. *Biological Bulletin*, 4:24-39.
73	Muller, Herbert J. 1922. Variation due to change in the individual gene. *The American Naturalist*, 56:32-50.
46	Riddle, Oscar. 1924. Any hereditary character and the kinds of things we need to know about it. *The American Naturalist*, LVIII:410-425.

Although the test site was not publicized, it was discovered by several other WWW sites, including the highly regarded MendelWeb site that provides access to Mendel's own work and other Mendeliana. The operator of that site has now added a cross-reference to *Classic Papers in Genetics*. The cross reference may be found at: http://netspace.org/MendelWeb/MWref.html. *Classic Papers in Genetics* has also been discovered and cited by the developers of the Natural History of Genes project (http://raven.umnh.utah.edu/umg/kits/kit.dna/online.html), which is an effort to develop a hands-on genetic science curriculum for middle and high school teachers.

The demonstration site has also been referenced in *The World-Wide Web Virtual Library, History of Science, Technology and Medicine*. A private page (http://www.duke.edu/~clc5/so.html) intended to offer reference materials for the Forest

Hills Central High School Science Olympiad team recommends the site, as does the "BioEd: Biological Sciences Education Resources" page at the University of Washington. The History of Genetics Web Pages, of the History and Philosophy of Science Program at the University of California, Davis, also recommend the site.

In short, demand for access to electronically reprinted papers in classical genetics appears high. If the project is supported, then after it becomes fully functional, with hundreds of papers available, we expect it to be extremely popular.

One might yet ask, however, *is there really a need for more than fifteen thousand pages of such material?* The answer is: Yes, absolutely. If anything, even more pages would be desirable.

The format chosen for republishing (similar to that of a page in 6"x9" book, see examples and a discussion in Appendices III – VI) is desirable for several reasons, but it does not have a particularly dense word count. In comparison, just one issue of the journal *Genetics* is equal to approximately 1500 pages in the chosen ESP format. Thus, the planned word-count output for the project is less than that found in a one-year's subscription to *Genetics*.

From this perspective, the question about the size and scope of the project might be rephrased, *Can the foundations for the entire field of classical genetics be adequately represented in less space than a one-year subscription to one journal?*

Although *all* of the basics for all of genetics cannot fit in this amount of space, we believe that republishing hundreds of works will provide a very strong presentation of the foundations of genetics. And, we will do everything possible to increase our efficiency so that substantially more pages might be produced without requiring any additional resources. We are hopeful that efficiency increases might lead to an improvement of up to 50%, but to be prudent we are basing our promised results on our current level of efficiency (see discussion below).

Can it be done, on time and within budget?

In a three–year period, we expect to publish at least fifteen thousand pages of material establishing the foundations of Classical Genetics. This goal is based on work-requirement estimates determined during the demonstration project.

Although considerable variation in work required occurs across different papers, depending upon the quality of the original, the complexity of the typesetting, the number of figures, etc., our preliminary work indicates that it should be possible for one desk-top-publishing professional, working alone, to achieve a sustainable output of 25 pages of final, typeset-quality output per day. This estimate represents an average across a representative sample of types of works and includes all scanning, optical character recognition, proof-reading, correcting, formatting, and desk-top publishing. Assuming 200 work days per year (i.e., 20 days lost to other activities, such as updating computers and software, consulting with editors, etc.), that results in 5,000 pages of finished output per year.[1]

[1] This estimate assumes that the staff person has access to a very fast computer and scanner, since the actual scanning of pages and running of the optical character recognition software is a key rate-limiting step.

Our estimates for the time required to develop introductory commentaries and to perform cataloging and classification are based on experience with the demonstration project and on industry standards. For example, the National Library of Medicine estimates that a Medline journal scanner can index a typical paper in approximately 30 minutes. On this basis, we conclude that the 10% effort of the PI and the 15% effort of the librarian will be sufficient to produce the metadata to accompany the papers.

For our software development, we intend to rely primarily on "glue programming" — the writing of small, specific programs to tie existing software (usually commercially available) together into a coherent system. We do not propose to develop a state-of-the-art new system for electronic publishing. Rather, we aim to deploy established, working software efficiently. If other research projects, with the specific aim of developing new methods for supporting electronic scientific publishing, happened to produce systems that might be useful to us, and if these could be obtained inexpensively, we would of course consider adopting them (for example, see Kirstein & Montasser-Kohsari, 1996).

Is it worth it?

Whenever support for a new information resource is proposed, one must ask whether it will likely return appropriate value for the investment made. For databases of research findings, where much of the value lies in the currency and completeness of the holdings, a major concern is, paradoxically, that the project might be a success and thus become so important to the community that it will require support in perpetuity to maintain its value. With agency budgets under tight constraints, this is a serious concern.

This proposed project, on the other hand, will involve the immediate creation of information resources that will have lasting value, whether or not funding beyond that requested here is ever obtained. All of the important papers republished in this project will become available to the scientific community in perpetuity, once they are created, with no continuing requirement for long–term support. We will make every effort to keep papers available on our server, even after the end of funding, and we would certainly work with others to ensure that the papers would remain available, even if our personal ability to maintain the system was, for some reason, compromised.

Given the increasing significance of the ethical, legal, and social implications of genome and genetic research for all citizens, the need to facilitate genetic understanding among students, teachers, researchers, and the general populace is great. The proposed project will offer resources and help to readers world wide. The system will be designed specifically for ease of access and use by large numbers of readers, and it will also be designed to facilitate the addition of value by interested third-party developers. The creation of this electronic library will offer great and lasting benefits to the educational, scientific, and lay communities.

Because the site will be especially useful to teachers, the leverage effect of the project will be high. The quality of training in genetics and biology might well improve world wide as a result of this project.

WORK PLAN

We will select and publish electronically a substantial body of classical genetics literature. The manuscripts will be prepared as type-set quality files in Adobe Acrobat

PDF format and distributed via a World Wide Web server. This will allow the electronic distribution of files that can either be read on-screen or printed with type-set quality, using freely available browser software.

Scope of Work: Type of Material to be Included

The works we republish will be primarily original and secondary literature, but we will also publish large excerpts from many selected monographs, such as Weismann's "The Germ Plasm Theory," Bateson's "Mendel's Principles of Heredity," Morgan's "The Theory of the Gene," etc.

Although we consider it unlikely that any one person would be interested in *all* of the works we will produce, we will nonetheless publish systematic, large holdings so that subsets of our output will fully meet the needs of many different users. Our plan is to have the site provide a substantial body of literature representing the foundations of classical genetics, not merely a few papers here or there labeled as classics themselves. We expect that users and third-party developers will add much value to the resources we create by assembling sets of our documents into larger, annotated resources.

We will include some papers that we now know to be simply wrong, but which represented important thinking at the time. For example, we will publish works on both sides of the biometrician-Mendelian debate from the early 1900s. And, we will present Castle's papers that argued against the notion of a simple linear genetic map for some time, and also those critical of the chromosome theory of sex determination.

Papers that are not now "classics" but which help capture the contemporary thinking and excitement of the time will also be included. For example, the demonstration site has recently added an 1899 paper by Bateson in which he lays out his notion of what sort of research will be required to begin the process of understanding heredity:

> What we first require is to know what happens when a variety is crossed with its nearest allies. If the result is to have a scientific value, it is almost absolutely necessary that the offspring of such crossing should then be examined statistically. It must be recorded how many of the offspring resembled each parent and how many shewed characters intermediate between those of the parents. If the parents differ in several characters, the offspring must be examined statistically, and marshalled, as it is called, in respect of each of those characters separately.

Those unfamiliar with Bateson's thinking before the rediscovery of Mendel may be surprised to find such a nearly a perfect prediction of Mendel's methods. It is no wonder that Bateson apparently became an immediate convert to Mendelism — he was already a Mendelian before he read Mendel's work.

It would also be interesting to publish works that demonstrate how prevailing social beliefs can find their way into the scientific literature. For example, one early single-author textbook, that went through four editions from the 1920's to the 1940's, showed a remarkable transformation. The first edition commented on on the "well-established" genetic inferiority of non-whites, but the fourth edition asserted that there was no scientific evidence for racial claims of genetic superiority.

Unlike some other belief systems, scientific belief is based on the concepts of skepticism and refutability. Therefore, we will also publish papers that demonstrate how scientists exhibit skepticism, yet can change their minds. For example, in the sample collection already published, there is an early paper from Morgan in which he expresses

real skepticism about the ease with which Mendelians make ad hoc modifications to their model to accommodate new findings:

> In the modern interpretation of Mendelism, facts are being transformed into factors at a rapid rate. If one factor will not explain the. facts, then two are invoked; if two prove insufficient, three will sometimes work out. The superior jugglery sometimes necessary to account for the result, may blind us, if taken too naively, to the common-place that the results are often so excellently explained because the explanation was invented to explain them. We work backwards from the facts to the factors, and then, presto! explain the facts by the very factors that we invented to account for them. I am not unappreciative of the distinct advantages that this method has in handling the facts. I realize how valuable it has been to us to be able to marshal our results under a few simple assumptions, yet I cannot but fear that we are rapidly developing a sort of Mendelian ritual by which to explain the extraordinary facts of alternative inheritance.

Yet just a short time after this paper, Morgan has become a world leader in establishing the Mendelian model of inheritance.

Scope of Work: Specific Material to be Included

An editorial board (see below) will assist in selecting works to be published. The detailed list of publications will evolve as the project unfolds, but some specifics may be given here. We will first select *potential* topic areas, such as Mendel's own works, Galton and the biometricians, the rediscovery of Mendel, physiological genetics: flower color, physiological genetics: inborn errors of metabolism, etc. Lists of planned topic areas (including sample bibliographies for each) will be published on the server and comments sought.

Our range of potential topic areas will span the field of classical genetics, which, by our loose definition, fundamentally spans the fields of genetics prior to the elucidation of DNA structure. Assuming appropriate permissions may be obtained, our set of publications will end with the topic area of the chemical gene, and will close with the Watson-Crick papers of 1953.

When a topic area is selected for publishing, a (partially) annotated bibliography will be developed and made available to users of the system, with comments again sought regarding the appropriateness of the material included, relative priorities, etc.

As topic areas are selected and input received from the user community, specific works will be selected for inclusion and placed into a scheduling queue and attempts will be initiated to obtain high-quality copies for input. The actual scheduling of publishing will depend in part upon the availability of adequate originals for input. However, we will strive to follow the priority recommendations of users and to achieve a balance in the material published. That is, initially we will try to publish a balanced set of papers across a number of topic areas, then fill in details within topic areas later.

During the first 18 months of the project, we will plan to restrict our publishing to works no longer covered by copyright — that is, to works greater than 75 years old. As the project becomes established, probably around the 12-month point, we will begin negotiating with copyright holders for access to materials still covered by copyright, and we will turn our attention heavily to these materials during the last 18 months of the project.

Preliminary interactions with permissions editors during the demonstration project led us to believe that the proposed project is sufficiently different from other publishing

efforts that permissions editors may be initially confused by requests and may therefore respond conservatively. However, we expect that as the project unfolds and begins to develop a following in the scientific and other communities, and as we have our technology in place and can document what we will be able to do to protect the rights of copyright holders, then obtaining permissions will be more straightforward.

Editorial Board

A key part of the proposal will involve the recruitment of editors whose primary responsibility will be to help select works for inclusion in the publishing project. Editors will also be encouraged to write introductory materials for some republished works (especially those in their area, or for papers they specifically recommended), but this is not a requirement for service as an editor. Assistance from editors in locating actual physical source materials will also be welcome.

Charles Laird and Steven Henikoff of the Fred Hutchinson Cancer Research Center have already agreed to serve on the board, as has Elliot Meyerowitz of Cal Tech, Jasper Rine of the University of California at Berkeley, and Lindley Darden of the University of Maryland. Other potential board members have been identified and will be recruited when the project receives support. We will also be open to suggestions for editorial board membership.

Project Deliverables

The project will deliver fifteen thousand pages of typeset-quality electronic reproductions of works that establish the foundations of classical genetics. These will be accessible through a World-Wide Web server that will offer both browse list interfaces as well as a variety of searchable interface systems. An annotated browse-list interface was used for the demonstration project and a printed copy of that interface is available in Appendix II.

The papers that we produce will be designed to be as attractive and easy to read as would be expected in a quality paper publication. Originally, we hoped to publish papers in a format that was a direct copy of the format of original publication (see Appendix III), but that proved too time consuming, so we now plan to use a standard format for all publications (Appendix IV).

Our standard format will be designed both to produce a useful, readable format for original works, but also to provide a basis for the easy development of annotated versions of papers. How this might be done is explained in Appendix V. An example — an annotated version of Mendel's paper — is presented in Appendix VI.

Project Evaluation

The sole purpose of this project will be to provide useful information and services to users. Thus, continuing evaluation of how well the project is meeting the needs of users will be essential. Because users will interact with the system in an on-line, real-time fashion, it will be possible to collect user satisfaction information on line. Opportunities to provide spontaneous feedback will be given to users throughout the pages of the interface. In addition, we may from time to time initiate a more structured approach to

soliciting feedback, perhaps by periodically offering the opportunity to sign a guest book, or to respond to specific questions.

At the project gets underway, we will take steps to publicize the project, and to call attention to the opportunities for user input, by sending press releases and other announcements to appropriate publications, such as *Genetics*, *Science*, *Nature*, *Nature Genetics*, etc. In addition to scientific journals, we will also publicize the project in the lay press and in media aimed at K-12 teachers.

IMPLEMENTATION ISSUES

Efficiency is the Goal

Our goal is to deliver the maximum amount of useful information in the most cost-effective, efficient manner possible. Our choice of products and services to deliver, as well as of technologies to use, will be driven by that ultimate goal. The scheduling of our work over the three years of the project will also be done with an eye on efficiency.

Time Line

We seek three years support for the project. Approximately three months will be spent in preparatory activities: recruiting of the desktop-publishing staff person, editorial board establishment, hardware/software acquisition, staff training, and developing and testing our methods for high-throughput production. In parallel, we will begin initial topic-list and publication selection, site design, document design, and optimization of the data-entry and document preparation process. At the completion of the preparatory activities, we will begin republication in earnest. This will be accompanied with initial efforts to publicize the project and to call attention to the role of users in helping define the project.

At the twelve-month mark, we will evaluate our results to date, then make another concerted effort to publicize the project. By approximately the fifteen-month mark, we will have several hundred documents available on line and would have a well-established electronic publishing activity under way. We should also have a substantial user base. At this time, we will begin preliminary interactions with other publishers, seeking permission to begin republishing works protected by copyright. With an established project underway, and with specific want lists of documents available, and with specific tools developed to support electronic documents distribution, we will then be able to make very clear and precise requests of publishers.

We expect that the obvious value of our work, combined with our willingness to work closely with publishers to maintain their rights to copyright works, and to publicize and help attract buyers for their products, will gain us reasonable arrangements with many publishers, especially those associated with scholarly societies.

At eighteen months, we expect to turn our attention heavily towards acquiring and republishing newer works (i.e., works from the 1925-1953 period).

Work Flow

Once we are in high-throughput production mode, our general pattern of work flow will be as shown in Figure 1. By establishing and enforcing the goal of full automation for activities below the dotted line, we hope to produce a system that is maximally

scalable, while also providing good support for independent, third-party individuals who might wish to add value to our project.

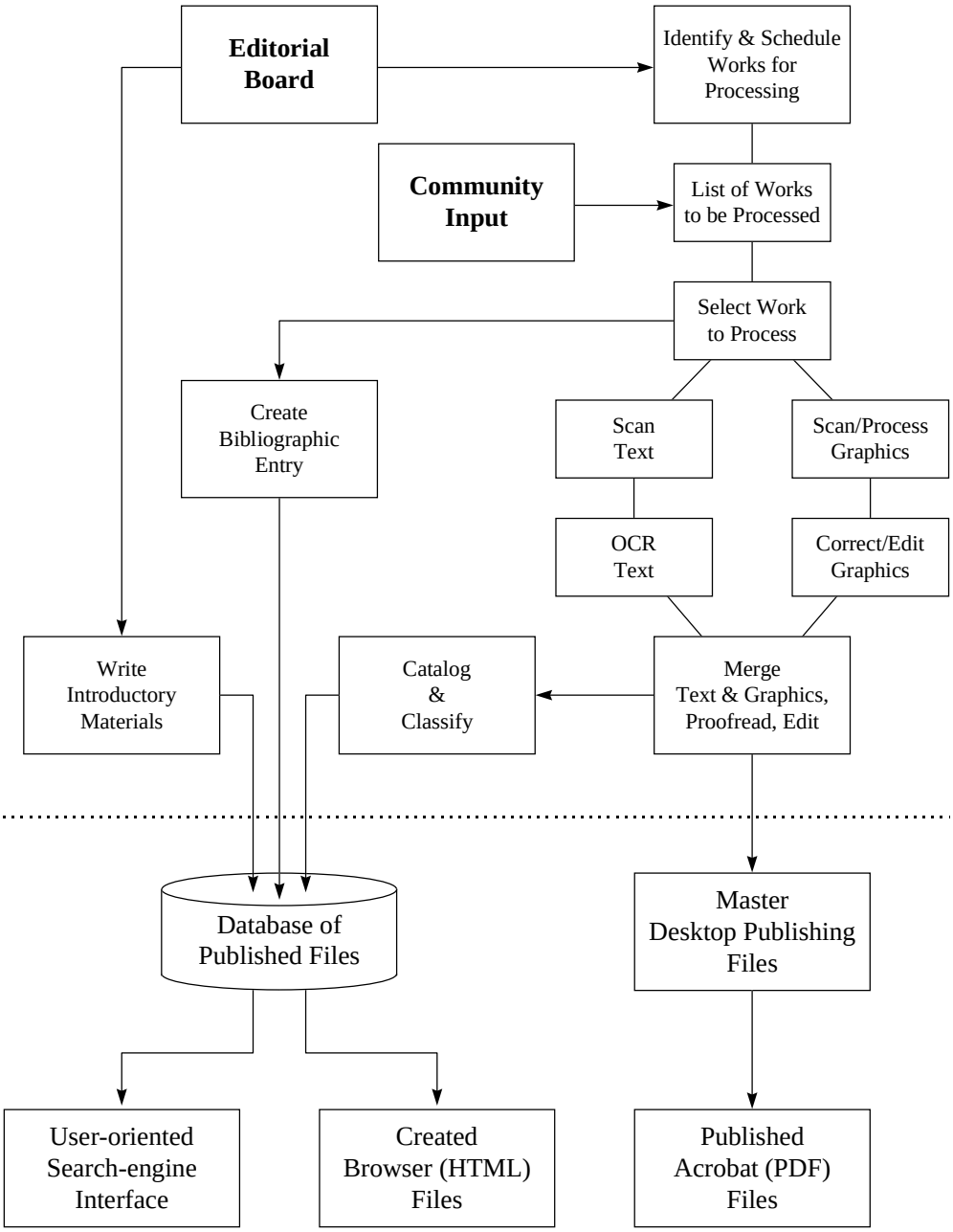

Figure 1. Activity flow in the production of electronically republished documents. To the extent possible, all activities below the dotted line will be fully automated.

We will constantly be looking for opportunities to move the "automation line" upward. At present there are no established tools that would let us efficiently manage standard files at a level higher than desktop-publishing format. However, for a variety of reasons it would be desirable for us to do so and we have noted with interest announcements from companies who allege that they will bring to market such products in the near future.

Staffing — Level of Effort

Preliminary experience with our demonstration project indicates that a skilled operator can produce more than 25 pages per day of finished, type-set-quality output. Preliminary experience also indicates that it might be possible to substantially improve that rate through the use of very high speed scanners and optical-character-recognition software and other improvements. However, to be prudent, we are basing our promised deliverables on a rate based upon actual experience.

We propose to hire one full-time staff person to scan, proof, type-set, and publish documents. During the preparatory phases, we will use contractors and professional trainers briefly to help establish the database system at the heart of the project and to optimize the scanning/publication process, since that will be the rate-limiting step for the entire effort.

The PI (Robbins) will devote 10% of his time to the project, and a librarian/cataloger will be used at 15% effort. These levels of effort have been estimated based on the preliminary data and on industry standards for literature cataloging and indexing.

Technical Issues

As noted earlier, our goal is to deploy efficiently established software and hardware tools, not to develop new systems for electronic publishing. We have chosen to use World Wide Web for distribution and Adobe Acrobat PDF files for content because these are now well established technologies and both have client software readily available for general use, on all platforms, either free or for very reasonable prices.

We chose Acrobat as our distribution file format because it allows complete typeset-quality control over formats, yet is already integrated into WWW technology, with improved integration to come. Although the format is proprietary, reader software is freely available on all major computer platforms.

We will have a few relatively small technical projects that we will need to complete to meet all of our goals. First, we will require a database for bibliographic, cataloging, and classification information. This will be used to manage our holdings and to generate automatically WWW browse lists in various formats. We have secured the services of Stan Letovsky to assist in this project. Dr. Letovsky's extensive experience in developing production-quality tools for other bioinformation resource projects convinces us that he will be able to develop a system that meets our needs, within the time and budget allocated. Much of this will be a port from other, working systems, and we do not expect and significant new development here.

One new tool that we will need is a method to allow the page-by-page merger of multiple PostScript files into a single file. This will be required for our planned support-for-annotation system (see Appendix V), and will have other uses as well, such as automatically adding registration marks or clipping guides, should one desire to use our files as camera-ready copy for printing.

Users searching our system will receive data structures that are essentially lists of bibliographic entries. We would like to develop a small Java tool, to be accessible via WWW, that can be used to display and manipulate such bibliographic lists through the visual metaphor of a drawer in a card catalog. The user will be able to move forward and backward through the "cards in the drawer" and will be able to make copies of selected

cards for later retrieval. This is a fairly modest project and is cleanly bounded — that is, its successful completion will add convenience and utility to our interface but any delay in getting it on-line will not significantly affect the delivery of our primary product, that is, the republished literature itself.

Much of our development work will be of a "glue programming" nature — that is, it will involve primarily the development of small bits of code that will serve to tie together other, larger components into our system. A glue programming approach will allow us to take maximum advantage of existing tools to support electronic publishing, while at the same time delivering a consistent, novel product to our users.

Other Issues

Cataloging and Classification

The value of such a republication effort will depend, in part, upon the ease with which readers can locate documents of interest. At the Fred Hutchinson Cancer Research Center we are just beginning to deploy World Wide Web publishing as a significant form of internal communication. As part of this effort, the FHCRC library will be charged with developing means for cataloguing and classifying (C&C) WWW publications, and the FHCRC WWW publishing program will deploy user tools to support C&C-based retrieval. We will be able to take advantage of these on-going activities for the *Classic Papers in Genetics* project without having to request extensive funds for their support.

Eve Ruff, the Director of the FHCRC library, also plans to seek external support for exploratory C&C work for electronic publishing and she has agreed to use the *Classic Papers in Genetics* project as a test bed for this effort. This will provide additional leverage for the project. Therefore, in this proposal we seek only sufficent support to allow for the actual cataloging and classification of our publications. We do not seek support for the development of tools or schemes for such cataloging and classification — that will be developed independently at the Center.

Design

To be effective, the site will have to be well designed and the documents themselves will have to be designed and typeset with an eye toward maximum usability. We have already begun experimenting with the type-set design for the republished documents. Some of these experimental designs may be seen in the documents currently retrievable from the test site, or as illustrated in Appendix IV.

The FHCRC has hired a consultant to assist us in developing our WWW systems as a form of electronic publishing and in developing appropriate visual metaphors to help users interact with our site effectively. Much that will be learned from that project will be freely available to the project proposed here.

Copyright Issues

The maximum copyright duration for works published in the United States prior to 1976 is seventy-five years. This means that all scientific literature first published in the United States prior to 1922 is now in the public domain. Since this spans the crucial beginnings of modern genetics, it will be possible to republish hundreds of valuable

papers, and even whole monographs, without having to seek permission from copyright holders. Preliminary discussions with the permissions offices of some key journals have been initiated regarding publications still covered by copyright. These will be extended once the project gets underway.

Our goal is to make our publications freely available to the user community. However, if publishers require that we place some access restrictions on copyrighted works, in order to be allowed to republish them, then of course we will accede to the requirements of the copyright holders.

We also plan to claim copyright in any new materials created for this site, including introductory materials, annotations, etc. Since we aspire to becoming self-supporting at some point in the future, we must claim copyright in our creations so that we can retain the right for any reuse in a commercial setting.

FUTURE WORK

We propose this project as a stand-alone effort that will continue to deliver lasting value after its completion. At the same time, we note that there is a tremendous potential for ongoing related projects. An obvious follow-on would be to publish papers that deal with the foundations of molecular genetics. We will develop our project in a way that facilitates and permits follow-on projects, but neither requires nor expects them.

We believe that this electronic publishing project does have a reasonable possibility of becoming self supporting, but, like the internet itself, will require some base support to get it established and to demonstrate its utility to the user communities. Many on-line publishing effort now try to supplement their income via advertising and some offer extended periods of free access before the institution of a proposed subscription fee. We will explore opportunities for becoming self supporting throughout the project. Possibilities that are most appealing would involve the notion of selling site subscriptions, so that all machines from a particular domain (e.g., FHCRC.ORG) would have unlimited access to the materials after some component at FHCRC.ORG (perhaps the library) paid a subscription fee.

To maintain the usefulness of the site for all potential users world-wide, we would probably move in this direction only if we could devise a way to provide general access to the materials to everyone for no fee, but offer additional premium services to those wishing to pay a fee. For example, in some of our preliminary discussions with publishers we found that they may be more comfortable in granting permissions to republish copyrighted materials if there will be some limits placed on the distribution of the electronic materials. Thus, if it were required in order to gain permission to republish some materials, we might institute a policy of providing access to abstracts of copyrighted material to all comers, but full-text access only to subscribers.

In any event, the goal of becoming self-supporting applies to the follow-on period after the end of support for the period requested and is described briefly here only to illustrate that we will be exploring ways to maintain the site without requiring long-term grant support from a federal agency.

SUMMARY

In this project, we propose to develop a system for the electronic republishing of scientific papers in a type-set-quality format. Preliminary results from a demonstration project have established the feasibility of such a project and have offered some indications of great need and demand for such a project.

We will republish works establishing the foundations of classical genetics and we will also provide a variety of tools to facilitate access and use of our publications. Our site will be designed to function in a federated information infrastructure and we will actively support the work of third-party, value-adding developers.

In 1900, genetics was just beginning as scientific discipline. As 2000 approaches, the Human Genome Project promises to deliver molecular-level information detailing our genetic heritage. Now, the early results of the HGP hold out great promise in areas from agriculture to biotechnology to medicine and beyond. As the HGP runs to completion, its effects will spread throughout society, stirring increasing interest among informed citizens. Understanding the significance of the HGP in part depends upon a basic understanding of genetic principles. This project will help bring that understanding to a wide audience.