Model Formulation ■

GeneClinics:

A Hybrid Text/Data Electronic Publishing Model Using XML Applied to Clinical Genetic Testing

Peter Tarczy-Hornoch, MD, Paul Shannon, Patty Baskin, MS, Miriam Espeseth, Roberta A. Pagon, MD

Abstract GeneClinics is an online genetic information resource consisting of descriptions of specific inherited disorders (''disease profiles'') as well as information on the role of genetic testing in the diagnosis, management, and genetic counseling of patients with these inherited conditions. GeneClinics is intended to promote the use of genetic services in medical care and personal decision making by providing health care practitioners and patients with information on genetic testing for specific inherited disorders. GeneClinics is implemented as an object-oriented database containing a combination of data and semistructured text that is rendered as HTML for publishing a given ''disease profile'' on the Web. Content is acquired from authors via templates, converted to an XML document reflecting the underlying database schema (with tagging of embedded data), and then loaded into the database and subjected to peer review. The initial implementation of a production system and the first phase of population of the GeneClinics database content are complete. Further expansion of the content to cover more disease, significant scaling up of rate of content creation, and evaluation redesign are under way. The ultimate goal is to have an entry in GeneClinics for each entry in the GeneTests directory of medical genetics laboratories—that is, for each disease for which clinical genetic testing is available.

■ **J Am Med Inform Assoc.** 2000;7:267–276.

Gene discoveries resulting from the Human Genome project can be translated into genetic tests that can improve medical care and expand personal choices for persons with inherited disorders.[1,2] Clinicians in domains ranging from pediatrics to neurology and oncology now need up-to-date, systematically organized information on the rapidly changing arena of genetic testing.[3–7] Existing online molecular genetic databases[8–11] designed for the research community do not contain information that can readily be used by clinicians. Furthermore, existing print and online resources do not provide practicing clinicians with a source of synthesized current information on genetic testing for diagnosis, management, and counseling that is applicable to their practices. Busy clinicians need to quickly find answers to such questions as ''Does genetic testing play a role in the diagnosis of neurofibromatosis?'' and ''How do I interpret a Huntington's test?'' Such an information source would need to be up to date, readily available and, ideally, integrated with the primary genomic databases used by the research community.

This paper describes the model developed by the GeneClinics project to electronically create, store, and distribute a Web-accessible information resource relating genetic testing to patient care. A goal is to merge the best attributes of traditional print publication (including book chapters and peer-reviewed articles), electronic publication, and online databases. The model is nearing full implementation, and evaluation processes are being designed. The schemas used to capture clinical and biological information consisting of both free text and discrete data elements are presented along with the use of XML and a hybrid

Affiliation of the authors: University of Washington, Seattle, Washington.

This work was supported by grant P41-LM06029 from the National Library of Medicine and the National Human Genome Research Institute.

Correspondence and reprints: Peter Tarczy-Hornoch, MD, Department of Pediatrics, Division of Neonatology, Box 356320, University of Washington, Seattle, WA 98195-6320;
e-mail: ⟨pth@u.washington.edu⟩.

Received for publication: 8/18/99; accepted for publication: 1/10/00.

text/data model to populate the schemas. A description and illustrative example outline the editorial process, information flow, and tools used to populate the database and render the content on the World Wide Web. The current state of the initial production model after population with content for 57 genetic diagnoses is discussed along with work planned.

Background

Electronic Genetic Information Resources

The genetics research community, focused on gene discovery, is ahead of the clinical community, focused on genetic testing, in the number and sophistication of electronic information resources and standardized data models available for its use. For researchers, information on DNA nucleotide sequences can be found in GenBank[9] and other sequence databases. Mutation information is contained in the Human Gene Mutation Database (HGMD)[10] and in locus-specific mutation databases such as the P53 database.[11] A number of databases, including Entrez,[9,12–14] contain information on protein sequences, structure, and function. These databases contain a rich set of data and sophisticated tools to search and analyze them.

In contrast, clinicians need information on genetic test availability and application. Genetic test availability can be determined using the GeneTests database and directory.[15] Clinicians seeking to apply genetic testing to clinical care have had less sophisticated tools at their disposal. The most heavily used electronic clinical genetic resource is Online Mendelian Inheritance in Man (OMIM),[8] which was developed in the mid-1960s from the personal notebooks of Victor McKusik. It is a catalog of human genes and descriptions of genetic disorders. OMIM has no underlying biological data model, nor is the curation of the contents broadly distributed. The OMIM entries are similar to those in an annotated bibliography and are generally organized chronologically, from early clinical descriptions through gene mapping efforts to gene discovery and cataloguing of allelic variants. Information sought by clinicians, such as diagnostic criteria, current status of genetic testing, and genetic counseling issues, are rarely discussed. Little attempt is made to synthesize information or even to delete inaccurate information. The OMIM content is intended to reflect the evolution of our understanding of the molecular genetic basis for phenotypes (in the diachronic model developed by McKusik). OMIM in not designed to provide clinicians with information that can be used quickly and conveniently in patient care.

Electronic Submission

Efforts to use electronic submission of data for publication have ranged from submission of documents in journal-specified format (e.g., the JAMIA process) to LaTeX and other templates and to customized tools. The benefit of templates and customized tools is that the submitted information can be interpreted and processed. The genomic community has had success curating many of their databases in a distributed fashion as researchers submit new data and annotate existing data using Web submission forms and uploadable templates. The highly structured nature of the data lend themselves well to this approach. The challenge is greater for clinical medicine, where the structure of the underlying data is more variable than in the more limited domain of genomics. The Cochrane Collaboration[16] has been able to create specialized tools to permit submission of both text and data in structured formats (e.g., the data for the meta-analysis and the accompanying text). The eMedicine group[17] has created a template-based electronic textbook publishing solution that allows authors and editors of the textbook to update their work via the Internet.

Model Description

Editorial Process

To provide clinicians with information focused on the use of genetic testing in patient care, the GeneClinics electronic publishing model adopted a distributed, expert-authored, peer-reviewed online publishing process subject to editorial oversight. The model accords with a number of guidelines and recommendations for the publication of quality information on the Internet, including those of the editors of the *Journal of the American Medical Association*,[18] the HONcode,[19] and the Six Senses Review[20] and criteria for assessing the quality of health information on the Internet.[21] The editorial team consists of an editor-in-chief, a managing editor, a librarian, associate editors, and an editorial board. The editors are recognized experts and opinion leaders in medical genetics who have particular expertise in clinical diagnosis, management, and molecular testing.

Diseases to be profiled are selected by an editorial board on the basis of clinical importance and the availability of genetic testing (determined from the GeneTests database[15] of available genetic tests, formerly Helix[22]). The authors and peer reviewers selected by the editors are clinicians and molecular pathologists or geneticists who are authorities in medical genetics and related fields. The authors have a renewable two-year term of authorship, during which

they are required to update their content based on changes in the use or availability of genetic testing.

Schemas

Two quasi-independent, interlinked schemata have been used in the GeneClinics database—the "profile schema" and the "biological schema." The profile schema is a way of organizing the bulk of the free text in the database. The biologic schema is a way of organizing the bulk of the discrete biological data in the database.

Profile Schema

The profile schema, readily apparent to any reader, has been refined through an iterative process and consists of the structure of the disease profile itself. The profile is organized (implicitly) as a set of answers to a series of questions that clinicians might ask about a disease from the perspective of genetic testing. The profile schema consists primarily of semi-structured free text with enough flexibility to accommodate unusual diseases but enough structure to enforce an appropriate amount of similarity across all disease profiles.

The key sections are Summary, Diagnosis, Clinical Description, Management, Genetic Counseling, Molecular Genetics, Resources, and Reference. Sections, in turn, have both mandatory and optional subsections. For example, the Diagnosis section includes mandatory subsections on clinical diagnosis and molecular diagnosis and an optional subsection on laboratory testing (e.g., nonmolecular genetic testing). Discrete data elements are embedded in some sections (using restricted controlled vocabularies where available). Some subsections (i.e, the Resources section and the Reference section) are essentially all data (as opposed to free text).

Biological Schema

The biological schema is effectively embedded as discrete data inside sections of the profile schema (largely in the Molecular Genetics section). The guiding principle for the biological schema was the creation of database entities that try to model, with all possible fidelity, actual entities in the world. The notions of "actual entities in the world" were drawn from an evaluation of the entities in the current GeneClinics disease profiles and from the advice of clinicians, bioinformaticians, and molecular geneticists. In addition, elements of a still-evolving shared ontology for molecular biology, which is being developed by the Molecular Biology Ontology Working Group (formed at the Intelligent Systems for Molecular Biology Annual Meeting in 1998) have been, and continue to be, incorporated into the GeneClinics data model (Figure 1).

Overall, the biological schema is intended to:

- Provide the organizational basis for the molecular genetics section of each profile
- Permit reuse of biologic entities (e.g., genes) across multiple phenotypes
- Permit formal specification of the molecular pathogenesis of diseases

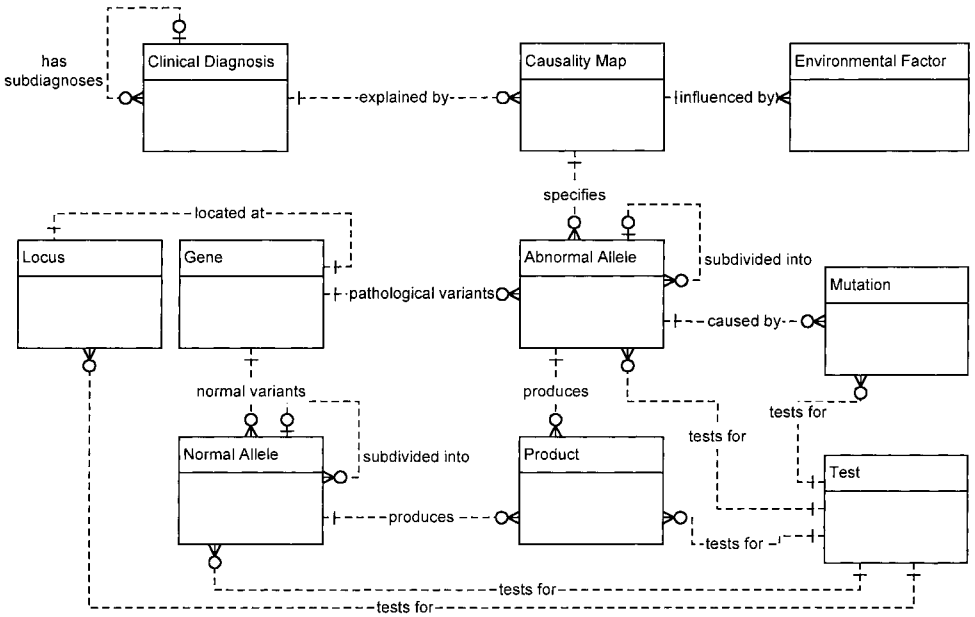

Figure 1 Top-level GeneClinics biological data model. This modified entity-relationship diagram illustrates the relationship between phenotype and genotype in the GeneClinics database as well as the relationship among the biological entities. This is a simplified high-level diagram of a subset of the biological data model, focusing on elements pertinent to the electronic publishing model. It does not represent the full entity-relationship diagram.

- Permit formal specification of genotype–phenotype correlations, including discrete data such as prevalence information for phenotypes and mutations
- Allow definition of the relationships among the steadily growing number of interrelated disease profiles (e.g., representing the fact that one phenotype has multiple independent genotypes)
- Provide the basis for bidirectional interaction with other genomic databases (aided by the use of international standard nomenclatures, as per HUGO, NCBI, and EBI recommendations)

Three key entities have been identified in the GeneClinics core data model: clinical diagnoses; pathologic allelic variants (i.e., disease-causing alleles with their related genomic entities of gene symbol, chromosomal locus, normal allelic variants, and gene products); and causality maps, which explain diagnoses in terms of pathologic allelic variants. As shown in Figure 2, the diagnoses are hierarchic—for example, the group of Charcot-Marie-Tooth (CMT) hereditary neuropathies contains a clinical subgroup, CMT1, which in turn as one variant, CMT1A, for which a specific causal mechanism is known. Similarly (but not shown in the figure), specific mutations are grouped into a set of alleles, which may be normal or abnormal. Associated with entities in the biological schema are discrete data elements—e.g., names for genes, loci, mutations, and products using controlled standard nomenclatures; disease prevalence; and testing sensitivity of specificity—as well as descriptive text, such as descriptions of normal gene products.

One important lesson that was learned from our experiences with GeneTests over the past six years and with GeneClinics over the past four years is that our data model must be flexible enough to accommodate the continual refinement of our understanding of the causal relationship between genotype and phenotype. Furthermore, clinical diagnosis and molecular diagnosis represent different approaches to medical practice, and they develop with distinctly different techniques, paces, and ontologies. For instance, the previously recognized single diagnosis of CMT was subdivided on the basis of phenotype into CMT1, CMT2, CMT3, CMT4, and CMTX. When the molecular genetic basis of these disorders was discovered,

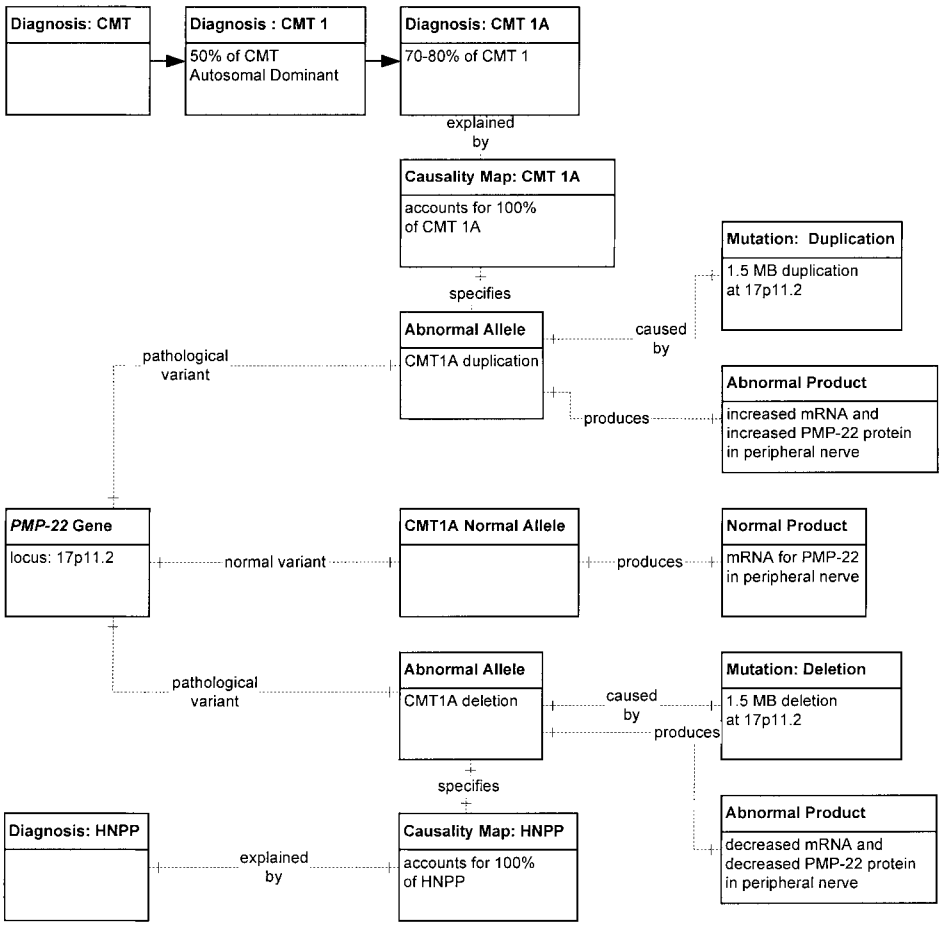

Figure 2 Sample instantiation of the GeneClinics biologic data model, showing the sharing of a single locus, gene, normal allele, and normal product by two distinct diseases caused by two different mutations (a deletion and a duplication of a 1.5-mega-base segment of DNA). The two diseases are Charcot-Marie-Tooth type 1A (CMT1A) and hereditary neuropathy with liability to pressure palsies (HNPP).

CMT3 was eliminated and CMT1 and CMT2 were further subdivided into CMT1A, CMT1B, CMT1C, CMT2A, etc. Only these last diagnoses show the classic correlation of a single disease with a single gene causality mapping (e.g., CMT1A syndrome is caused by duplication of the PMP-22 gene). The model also permits capturing of the fact that a different (deletion) mutation of the same gene is causal in hereditary neuropathy with liability to pressure palsies (HNPP).

Causality maps (Figure 1) are intended to capture our best notions of causation of both simple diseases, with a one-to-one relationship between gene and phenotype (e.g., NF1, related to neurofibromatosis type 1), and complex diseases with multiple genes and identical or overlapping phenotypes (e.g., CMT). The model has so far permitted the representation of multiple types of complex phenotypes (such as Alzheimer disease), in which mutations in a single gene can cause a phenotype (e.g., early-onset familial Alzheimer disease) and multiple genes interacting in a polygenic manner can cause the same phenotype. These maps could be extended, as our knowledge grows, to include other, as yet unknown factors that play a role in the development of genetic diseases.

XML-based Hybrid Text-and-Data Model

From the standpoint of the authors, readers, and editorial staff, a particular GeneClinics entry contains the information for both schemata displayed in an integrated fashion. The biological schema serves as the underpinning for the profile schema subsections on molecular genetic testing, genetically related disorders, genotype/phenotype correlation, and prevalence. The molecular genetics profile section contains key biological schema information about genes, loci, alleles, and gene products involved in a particular disease as well as linkages to other genomic databases (including OMIM,[8] the Human Genome Mutation Database,[23] and the component databases in Entrez[12]). During the design phase of the GeneClinics model, the editorial staff expressed a desire to populate both the biological and the profile schemata using a single tool and a familiar document-based paradigm. They were opposed to a separate tool for maintenance and viewing of the biological schema and wanted tight integration with the text of the profile schema.

The GeneClinics project therefore developed a model in which both profile and biologic free text and discrete data are marked up in XML in a single document then parsed and stored in a single database. The GeneClinics DTD (document type definition) represents both the profile schema and the biological schema as well as some basic formatting and editorial markup elements. Figure 3 shows selected fragments of the current GeneClinics XML DTD. The first entity represents an instance of a free text entity. Notice that it permits text and also basic formatting and editorial markup as well as data in the form of clinical and biological entities. The first element reflects the top-level element of the profile schema. The second element represents a simplified version of the causality map from the biological schema (Figures 1 and 2). The DTD mirrors our ObjectStore database schema, which permits relatively straightforward interconversion of database objects and XML documents. Since the editorial staff works with the DTD, it also serves as a lingua franca (or at least a patois) that can be used and understood by both the editorial staff and the technical staff.

```
<!-- Entity for generic block of free text in-line -->
<!ENTITY  % inlineText
   "#PCDATA | emph | italics | newline | space | link | linkDest |
    cite | editorSignature | symbol | subscript | superscript |
    startEdit | endEdit | strike |
    %biologicalEntities; | %clinicalEntities;">

<!-- Top level of the Profile Schema -->
<!ELEMENT dzProfile                  (title,
                                      synonyms,
                                      authors,
                                      lastUpdate,
                                      summary,
                                      mainDiagnosis,
                                      clinicalDescription,
                                      differentialDiagnosis,
                                      management,
                                      geneticCounseling,
                                      molecularGenetics,
                                      resources,
                                      references)>

<!-- XML representation of Causality Map from Biological Schema -->
<!ELEMENT causalityMap           (briefDiscussion?, fullDiscussion?,
                                  relativePrevalence,
                                  inheritance*,
                                  abnormalAlleleOBJ*,
                                  epigeneticFactor*)>

<!-- XML representation of the Abnormal Allele from Biol. Schema -->
<!ELEMENT abnormalAlleleOBJ      (briefDiscussion?, fullDiscussion?,
                                  geneProductOBJ*,
                                  genomicDatabaseReferences?,
                                  mutationOBJ*)>

<!ATTLIST abnormalAlleleOBJ gene CDATA #REQUIRED>
```

Figure 3 Fragments of the GeneClinics XML document type definition.

Information Flow

The following are the steps in the creation, markup, storage, review, editing and display of a GeneClinics disease profile (Figure 4). With the exception of the XML editor, the tools used are written in C, perl, and Java and run on a Sun Enterprise 450 server running Apache under Solaris 5.6.

Step 1: Authoring with smart template. The expert authors use "smart" templates designed by the GeneClinics staff for their word processor (Word or WordPerfect) to create the database entries. It is a "smart" template in that context-sensitive author instructions are provided, authors cannot delete or re-

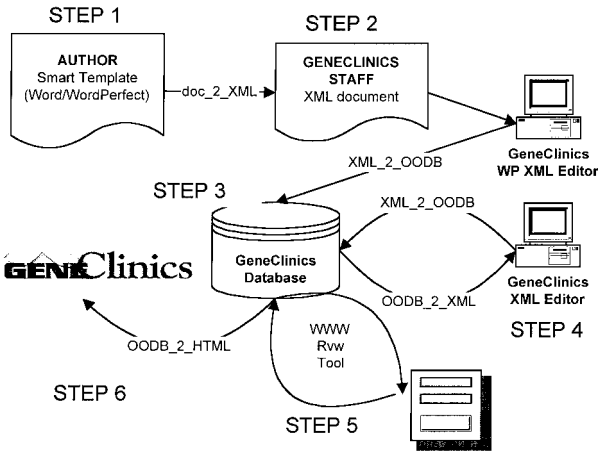

Figure 4 Overview of information flow in the GeneClinics electronic publishing model. This diagram shows the flow of information from the authors to the GeneClinics editorial staff and reviewers and, ultimately, to the readership. It also illustrates the necessary data format conversion.

arrange sections or subsections, and support is provided for some of the discrete data elements. The template mirrors the profile schema and represents part of the biological schema.

Step 2: Converting author document to XML. The doc_2_XML software parses the completed template and adds a) all the XML tags that can be inferred from the template, b) all elements of the profile schema, and c) some elements of the biological schema, both as discrete data and as free text. The joint profile/biological schema DTD is used. The staff use an XML editor to tag additional discrete data elements embedded in free text that are part of the biological schema (including tags and attributes that specify relationships between entities). In addition, the staff populate certain elements of the schema (additional resources, canned searches, and database linkages). The XML editor is built on top of WordPerfect SGML functionality using more than a dozen macros developed specifically for XML and GeneClinics (Figure 5).

Step 3: Loading XML version into object-oriented database. During the loading process, software parses the XML document into the various entities (free text, discrete data, relationships) that it represents and stores in an ObjectStore database (chosen for its native Java interface, robustness, and tolerance of schema evolution). As part of this step, linkages to other databases are parsed and verified and the appropriate foreign keys are stored (e.g., bibliographic entries are extracted and translated into PubMed identification numbers).

Step 4: Downloading XML version from object-oriented database into XML editor for iterative edits. As the disease profile is subject to the editorial process, the document passes through multiple revisions that are created in the XML editor and stored in the database, from which any revision can be retrieved on demand. This same tool is used for periodic revisions and updates of database entries.

Step 5: Uploading XML version from object-oriented data-

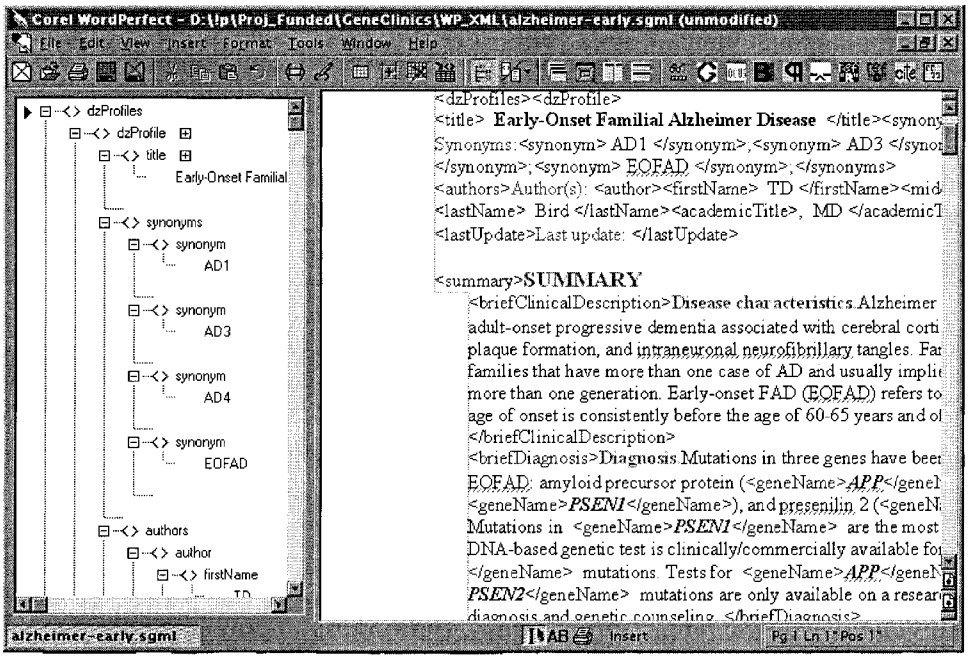

Figure 5 Screen shot of the WordPerfect-based XML editor open to early onset Alzheimer disease.

base into review tool for comments. The Java-based review tool (with appropriate authorization or authentication) operates against the database to permit the reviewer's comments to be tied to specific database elements and to render the contents of the database into various views, depending on the preference of the reviewer.

Step 6: Displaying disease profile as HTML document on the Web. Readers view the database content as HTML documents (Figure 6)

Presentation

The documents on the Web can be rendered into HTML from the contents of the database or from the XML document corresponding to the content of a given profile. The code that renders the HTML uses the data model specified by the XML DTD to determine order and hierarchy for display of data elements. The appearance on the screen (formatting) of classes of elements in the DTD is hard coded in the rendering code. For a selected disease profile the combined profile or biological data model is used to traverse all the children of that disease profile (e.g., all the elements of the profile and biological schema that belong in that profile). The GeneClinics project adopted the following criteria of Web site HTML design: low byte count to minimize download time for low-bandwidth users, cross-browser/cross-platform compatibility/readability, accessibility by lowest-level available browser (i.e., IE3, Netscape 3), maintainability, and access to a "printable copy" to avoid the difficulty of printing with frames.

Example

The Molecular Genetics section of the disease profile for achondroplasia can be used to illustrate the components of the model. The author completes the Molecular Genetics section of the word-processing template, first listing for each involved gene (in this case, just FGFR3) its symbol and name, its locus, a description of the normal allelic variants, and the name and a description of the normal gene product. For each phenotypic variant (in this case, just achondroplasia, but it could be CMT1A, -1B, -1C, or -1D), the author provides a name for the phenotype, its relative prevalence (in this case 100 percent), the symbols for the genes involved (in this case, FGFR3), a text description of the roles of the genes, a text description of the abnormal allelic variants involved, and a text description of the abnormal products, if known.

The populated template fields are converted to XML tagged entries corresponding to both elements of the schema profile (e.g., ⟨molecularGenetics⟩ tag) and to some of the elements of the biological profile (e.g., ⟨geneSymbol⟩ tag). This conversion makes it possible to identify and tag for one profile (Figure 1) a list of diagnoses (phenotypes) and a list of one or more sets gene, locus, product, and normal allelic variants—in this case, achondroplasia and the set comprising the FGFR3 gene, FGFR3 protein and its description, 4p16 locus, and a description of the normal allelic variants of FGFR3. At the end of the Molecular Genetics section of the XML document, the editorial staff manually inserts appropriate tags describing the causality after interpreting (with help from the author, if nec-

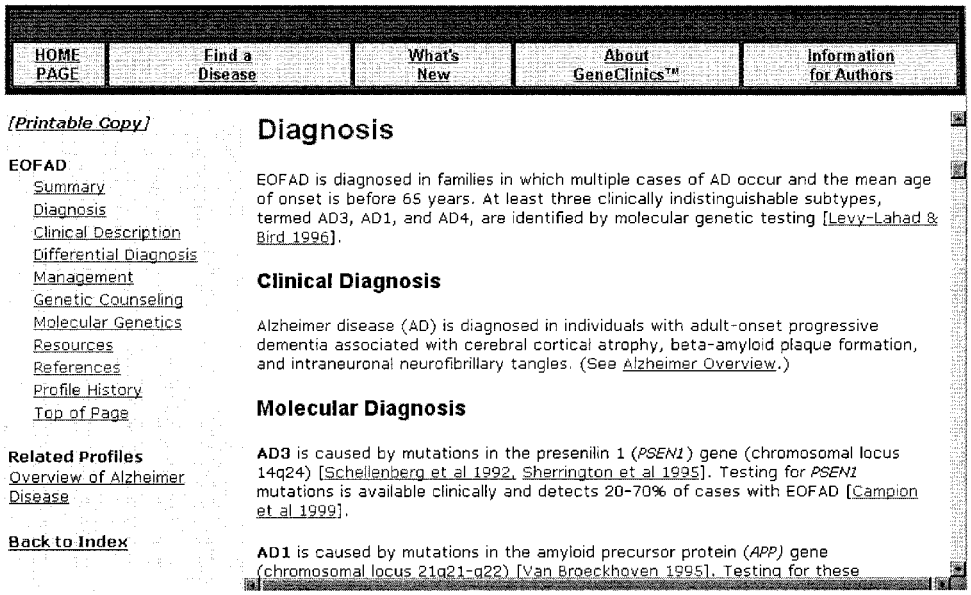

Figure 6 Screen shot of the HTML-rendered GeneClinics content for early onset Alzheimer disease.

essary) the text descriptions of the role of the genes involved, the abnormal allelic variants, and the abnormal products. In this case, the tags specify that 99 percent of the diagnosis achondroplasia is explained by mutations in the FGFR3 gene. The description of the mutations and the abnormal product remains the free text provided by the author. In this fashion, both the profile and biological components of the schema are populated using XML. The XML document is then stored in the database (roughly, each XML tag corresponds to a class in the database). The Molecular Genetic section can then be displayed from the database as HTML. First, using information from the biological schema, a table is generated displaying the genes, loci, and products involved. Then, for each phenotypic variant, a list of one or more causality maps is displayed as sets of gene, description of abnormal allelic variant, and description of abnormal product. Finally, after all the phenotypes and their variants have been listed, the description of the normal allelic variant and normal product of each involved gene is displayed.

Implementation

Work on the GeneClinics electronic publishing model was begun in 1995 and has been funded since 1997. We have only recently completed work on the initial production system, which is why the number of entries to date is relatively small. Because of the long lead time necessary to commission, obtain, review, edit, and publish the GeneClinics content, an intermediate electronic publishing model (populated profile schemas semi-manually converted to static HTML pages) was used initially.

The GeneClinics electronic publishing system is in transition from this interim model to the model outlined in this paper. The system comprising the template, XML editor, and database described above is in production and is being used for all new profiles. The review tool (Figure 4) is in limited use until the next generation of Web browsers are in general release, at which time direct manipulation of text in a browser window will be possible. The HTML rendering is currently (as of January 2000) being done nightly on a batch basis, and the desired state is generation "on the fly" from the database. Approximately half the profiles published under the interim model have been semi-automatically converted to XML and stored in the object database. We are in the process of converting the remainder of the profiles from the interim model to the object database model. There are currently 59 different published disease profiles (including both current and interim model entries), 21 profiles under review, and 37 profiles commissioned but not yet received. Usage is growing (currently between 400 and 700 substantive hits a day), although the site has not been widely publicized. A substantive hit is defined as access by a user to a specific disease entry in GeneClinics.

We are using the biological schema elements of the database to explore creation and maintenance of bidirectional links to other genomic and clinical resources. We are collaborating with the National Center for Biotechnology Information (NCBI) to establish and automatically maintain bidirectional links between GeneClinics and Entrez databases (e.g., PubMed, OMIM, LocusLink, and Genes and Diseases). We link *to* PubMed via PubMed identification numbers and plan to expand linkages to Entrez by using locus, gene symbol, and protein product as search keys. As part of the collaboration we have established links *from* relevant PubMed entries to corresponding GeneClinics entries using the NCBI "LinkOut" mechanism and the automatic regular publishing of GeneClinics "holdings files" to NCBI. We establish links to LocusLink using OMIM numbers as search keys. Following receipt of automatically electronically submitted pairs of OMIM numbers and disease names in a standard format, NCBI establishes links from LocusLink to GeneClinics. We are expanding this to bidirectional links with OMIM. We accept links from other databases via an external resource access servlet that queries our database by a variety of discrete data elements (including OMIM number, PubMed identification, and disease names).

Discussion

Electronic Genetic Information Resource

The GeneClinics editorial process (national expert authors, external peer review) is designed to address concerns about the accuracy and currency of an internally authored, internally reviewed clinical genetic resource like OMIM. Although not part of the electronic publishing model per se, this aspect has been important to readers and contributors to the database. The GeneClinics content and resultant profile schema were intended to systematically address a need, not met by the OMIM database, for directly applicable clinical information such as diagnostic criteria and the application of genetic testing to patient care. The structure of the profile schema is such that this information need should be met (since each major section of the profile schema is intended to answer a clinically relevant question), but this will need to be explicitly evaluated.

The biological schema underlying the GeneClinics model is less sophisticated than such schemas in the

genomic community but more than that of OMIM. The benefits of the biological schema for the GeneClinics project have thus far been the ability to readily interface with other genomic databases, the organization of the molecular genetics section, explicit representation of causality as a relationship, and the provision of a systematic framework for the discussion of genetic testing. As shown in Figure 1, tests apply to biologic entities which are linked to phenotypes; therefore, understanding and explicitly stating the causality between genotype and phenotype is a critical first step toward understanding and discussing testing. Other potential benefits, such as reuse of biologic entities, need to be evaluated as we scale up the database.

Electronic Submission

For authors, the GeneClinics model adopted a simpler template-based electronic submission model rather than a customized tool such as those developed by the Cochrane Collaboration or the eMedicine group. The completely nonstructured submission model was felt to not achieve the goals of having a profile schema that addresses specific questions relevant to clinicians. The Web tool-based model of the genomic databases was tried on a pilot basis, but the burden of the slow response time and primitive user interface on the Web led us to abandon this approach. The postsubmission editorial staff markup of discrete data elements (Step 2 in Figure 4) was adopted to achieve some of the benefits of a more highly structured tool that incorporates both text and data (such as the Cochrane tool or the genomic database annotation tools). How this process scales remains to be evaluated as we recruit more authors, start revising existing entries, and tackle more complex genotypic and phenotypic relationships.

Limitations

The most significant limitation identified thus far is that creating the infrastructure necessary for successful implementation of this model requires considerable expertise in publishing, editing, software, and database design. This limits the generalizability of the solution despite the theoretic portability of the design and approach. A related limitation is the relative immaturity of the tools available to create the infrastructure. The WordPerfect-based XML editor we developed is acceptable but not optimal. Object database technology is maturing rapidly, but many of the products are not yet ready for large-scale production use. Another limitation is the fact that the model assumes a consistent structure across documents (entries) in the database. This is especially true at the biological schema level, where most of the discrete data reside. The approach (in terms of capturing data) would lend itself much less well to a publishing environment in which the data were heterogeneous (e.g., a general medical journal publishing multiple kinds of clinical trials, reviews, and case reports).

Implementation Directions

We plan to refine our existing tools based on feedback from the editorial staff. We plan to continue to expand content by recruiting more authors. The ultimate goal is to have a complete entry in GeneClinics for each disease for which genetic testing is available (i.e., for all entries in the GeneTests database). We eventually plan to explore integration of clinical decision support tools driven by the biological schema. (A pilot project using belief networks for Huntington disease is underway.)

Evaluation

The plan is to evaluate both the process and the content. As currently planned, the process will be evaluated using a number of metrics in the context of scaling up disease profile production, which will test both the tools and the data model through both higher volume and greater heterogeneity of content. The evaluation of the content will be performed in collaboration with evaluators in the Department of Medical Education at the University of Washington. As planned, it will include in-depth evaluation of content accuracy, utility, and relevance using a selected sample of users (including domain experts, genetics professionals, and other clinicians) as well as polling all the users of the site as to their perceptions of quality and utility.

The authors thank the members of the GeneClinics informatics group, editorial staff, associate editors, editorial board, advisory group and, most important, our expert authors and reviewers. (See ''About GeneClinics'' at www.geneclinics.org for details.)

References ■

1. Fink L, Collins F. The Human Genome Project: view from the National Institutes of Health. J Am Med Womens Assoc. 1997;52:4–7.
2. Collins F. Medical and societal consequences of the Human Genome Project. N Engl J Med. 1999;341(1):28–37.
3. Andrews L. Social, legal, and ethical implications of genetic testing. In: Fullarton J, Holtzman H, Motulsky AE (eds). Assessing Genetic Risks: Implications for Health and Social Policy. Washington, DC: National Academy Press, 1994.
4. Statement of the American Society of Clinical Oncology: genetic testing for cancer susceptibility. J Clin Oncol. 1996;14: 1730–6.

5. Bird T, Bennett R. Why do DNA testing? Practical and ethical implications of new neurogenetics tests. Ann Neurol. 1995;38:141–6.
6. Clare AC. The genetic testing of children: statement of the Working Party of the Clinical Genetics Society (U.K.). J Med Genet. 1994;31:785–97.
7. Cotton P. Medical news and perspectives commentary: prognosis, diagnosis, or who knows? Time to learn what gene tests mean. JAMA. 1995;273:93–4.
8. National Center for Biotechnology Information. OMIM™: Online Mendelian Inheritance in Man home page. Available at: http://www.ncbi.nlm.nih.gov/Omim/. Accessed Jul 26, 1999.
9. Benson D, Boguski M, Lipman D, et al. GenBank. Nucleic Acid Res. 1999;27:12–7.
10. Cooper D, Ball E, Krawczak M. The human gene mutation database. Nucleic Acids Res. 1998;26:285–7.
11. Sedlacek Z, Kodet R, Poustka A, Goetz P. A database of germline p53 mutations in cancer-prone families. Nucleic Acids Res. 1998;26:214–5.
12. National Center for Biotechnology Information. Entrez Browser home page. Available at: http://www.ncbi.nlm.nih.gov/Entrez/. Accessed Jul 26, 1999.
13. Ostell J, Kans J. The NCBI data model. In: Ouellette B (ed). Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. New York: Wiley-Liss, 1998:121–44.
14. Pearson P, Francomano C, Foster P, Bocchini C, Li P, McKusick V. The status of online Mendelian inheritance in man (OMIM) medio 1994. Nucleic Acids Res. 1994;22(17):3470–3.
15. GeneTests Web site. Available at: http://www.genetests.org. Accessed Jul 26, 1999.
16. Bero L, Rennie D. The Cochrane Collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. JAMA. 1995;274:1935–8.
17. Medical textbooks for health professionals. eMedicine Web Site. Available at: http://www.emedicine.com. Accessed Jan 3, 2000.
18. Silberg W, Lundberg G, Musacchio R. Assessing, controlling, and assuring the quality of medical information on the Internet. JAMA. 1997;277(15):1244–5.
19. Health on the Net Foundation. HON Code of Conduct (HONcode) for medical and health Web sites. HONcode Web site. Available at: http://www.hon.ch/HONcode/Conduct.html. Accessed Jul 26, 1999.
20. The Six Senses Review: A Healthcare and Medical Website Review Program. FAQs. Available at: http://www.sixsenses.com/FAQ.html. Accessed Jul 26, 1999.
21. Health Information Technology Institute. Criteria for assessing the quality of health information on the Internet [white paper working draft]. Oct 14, 1997. Mitretek Systems Web site. Available at: http://hitiweb.mitretek.org/docs/criteria.html. Accessed Jul 26, 1999.
22. Tarczy-Hornoch P, Covington M, Edward J, et al. Creation and maintenance of Helix, a Web-based database of medical genetics laboratories, to serve the needs of the genetics community. Proc AMIA Annu Symp. 1998:341–5.
23. University of Wales College of Medicine. Human Gene Mutation Database at the Institute of Medical Genetics in Cardiff (U.K.). Available at: http://www.uwcm.ac.uk/uwcm/mg/hgmd0.html. Accessed Jul 26, 1999.