Authors:	Väliaho Jouni, Riikonen Pentti, Vihinen Mauno
Name of article:	Distribution of immunodeficiency fact files with XML – from Web to WAP
Year of publication:	2005
Name of journal:	BMC Medical Informatics and Decision Making
Volume:	5
Number of issue:	21
Pages:	1-11
ISSN:	1472-6947
Discipline:	Medical and Health sciences / Medical biotechnology
Language:	en
School/Other Unit:	Institute of Biomedical Technology

URL: http://www.biomedcentral.com/1472-6947/5/21
URN: http://urn.fi/urn:nbn:uta-3-643
DOI: http://dx.doi.org/10.1186/1472-6947-5-21

BMC Medical Informatics and Decision Making

BioMed Central

Research article

Open Access

Distribution of immunodeficiency fact files with XML – from Web to WAP

Jouni Väliaho*[1], Pentti Riikonen[1,2] and Mauno Vihinen[1,3]

Address: [1]Institute of Medical Technology, FI-33014 University of Tampere, Finland, [2]Department of Information Technology, University of Turku, FI-20520 Turku, Finland and [3]Research Unit, Tampere University Hospital, FI-33520 Tampere, Finland

Email: Jouni Väliaho* - jouni.valiaho@uta.fi; Pentti Riikonen - pentti.riikonen@it.utu.fi; Mauno Vihinen - mauno.vihinen@uta.fi

* Corresponding author

Published: 26 June 2005

BMC Medical Informatics and Decision Making 2005, **5**:21 doi:10.1186/1472-6947-5-21

Received: 11 March 2004
Accepted: 26 June 2005

This article is available from: http://www.biomedcentral.com/1472-6947/5/21

Abstract

Background: Although biomedical information is growing rapidly, it is difficult to find and retrieve validated data especially for rare hereditary diseases. There is an increased need for services capable of integrating and validating information as well as proving it in a logically organized structure. A XML-based language enables creation of open source databases for storage, maintenance and delivery for different platforms.

Methods: Here we present a new data model called fact file and an XML-based specification Inherited Disease Markup Language (IDML), that were developed to facilitate disease information integration, storage and exchange. The data model was applied to primary immunodeficiencies, but it can be used for any hereditary disease. Fact files integrate biomedical, genetic and clinical information related to hereditary diseases.

Results: IDML and fact files were used to build a comprehensive Web and WAP accessible knowledge base ImmunoDeficiency Resource (IDR) available at http://bioinf.uta.fi/idr/. A fact file is a user oriented user interface, which serves as a starting point to explore information on hereditary diseases.

Conclusion: The IDML enables the seamless integration and presentation of genetic and disease information resources in the Internet. IDML can be used to build information services for all kinds of inherited diseases. The open source specification and related programs are available at http://bioinf.uta.fi/idml/.

Background

Biomedical information is often very complex. Deciphering the roles of genes in human health and disease is a grand challenge for many reasons, including impediments to defining phenotypes, difficulties in identifying and quantifying environmental effects, technical problems in generating genotypic information, and the difficulties of studying humans [1]. The completion of the draft sequence of the human genome [2,3] and advances in molecular biology provide new opportunities to increase our understanding of the role of genetic factors in human health and disease [1]. The number of identified genetic diseases has increased exponentially [4]. The new knowledge can be applied to the prevention, diagnosis and treatment of diseases. This far, the knowledge of genetics has had a large role in the health care of only a

few patients and a small role in the health care of many [5]. The biomedical informatics holds great promise for developing informatics methods that will be crucial in the development of genomic medicine [6].

Most hereditary diseases are rare and the diagnosed patients for a condition are often randomly spread out in the world. One doctor usually has only a few patients with a disease. It is often difficult to find comprehensive and validated biomedical information related to rare diseases. In addition, it is more and more difficult to publish results in scientific journals only from a few cases even when they are interesting [7]. Still, all these pieces of information can contain clues to understanding the fundamental defects at molecular level and can help to develop targeted treatments. The scattering of the disease-related information to literature and Internet is a big obstacle especially for those interested in rare diseases. First of all, there may not be that much data for these diseases and secondly it may be very difficult to find and collect. Further, the user has often difficulties in assessing the quality of data.

There is an increasing need for tools and services capable of integrating information from a variety of sources. Clinicians and researchers could benefit from a more consolidated and unified view of the available biomedical data. Systems biology researchers need to integrate disparate information from multiple public sources to merge with their own experimental data to generate models of processes. Biomedical data mining attempts to extract information from biomedical databases by using e.g. automated natural language processing (NLP) techniques [8]. Processing of biomedical texts presents many challenges such as in the areas of terminology or ontology building, information extraction from texts, knowledge discovery from collections of documents, as well as sharing and integrating knowledge from factual and textual data bases, semantic annotation, etc. Without standardized nomenclature the information extraction (IE) about a particular subject from various resources is difficult. Due to ambiguity of terms, a search for a particular term often retrieves results for unrelated entities. Since there are also some technical problems arising from the diversity of computer hardware and software, there is a need for such a data form, that can be handled by any computer and which can be easily presented on any platform.

The Extensible Markup Language (XML) is a standard created by the World Wide Web Consortium (W3C) for characterizing the content and structure of documents [9]. It is designed to improve the functionality of the Web by enabling more flexible and adaptable information identification and presentation. XML allows to define tags and document structures for own context-specific use. It was derived from SGML (Standard Generalized Markup Language), the international standard for defining descriptions of the structure and content of different types of electronic documents [10]. XML is simpler than SGML, but it allows the use of richly structured documents over the Internet. Information encoded in XML is easy to read and understand, and easy to process by computers. In XML files, structured data are bounded by tags and attributes. XML tags, attributes and element structure provide context information that facilitates the interpretation of the meaning of content, thereby making it feasible to develop efficient search engines and agents and perform intelligent data mining, etc. The XML allows the separation of content, logic and presentation.

Beyond XML there are a number of additional specifications such as Document Object Model (DOM) [11], XML Schemas [12], XSL Transformations [13], and Resource Description Framework (RDF) [14]. XML will have a big role in integration and interoperation of biological databases. Some biomedical information models have been implemented using XML specifications [15,16], many of them being clinical models for electronic healthcare documents [17-19].

A unified data format of resources is required for comparison between similar diseases and reutilization of information. Here we present a new data model called fact file, which integrates biomedical information related to hereditary diseases into a Web and WAP accessible knowledge base. Our scope is wider than e.g. in gene oriented knowledge bases such as GeneCards [20], UniGene [21], or LocusLink [22]. The disease information sources are even more diverse than those for genetic information. The fact files concentrate on sharing and integrating biomedical knowledge from different sources. The presented data model can be applied to any hereditary disease.

The fact files were applied to build a comprehensive, validated knowledge base for primary immunodeficiencies (PIDs) called ImmunoDeficiency Resource (IDR) [23,24]. It is designed for different user groups such as researchers, physicians and nurses as well as patients and their families and the general public. The IDR is the major information source to immunodeficiencies in the Web. Fact files serve as the core of the IDR knowledge base.

Methods

The fact file data model and the Inherited Disease Markup Language (IDML) were developed to facilitate disease information integration, storage and exchange in the first place for immunodeficiencies, but in principle for any hereditary disease. The IDML is an XML specification and container for bioinformatical data on hereditary diseases. The fact file data model schema was defined according to W3C XML specification [12, see Additional files 1, 2, 3].

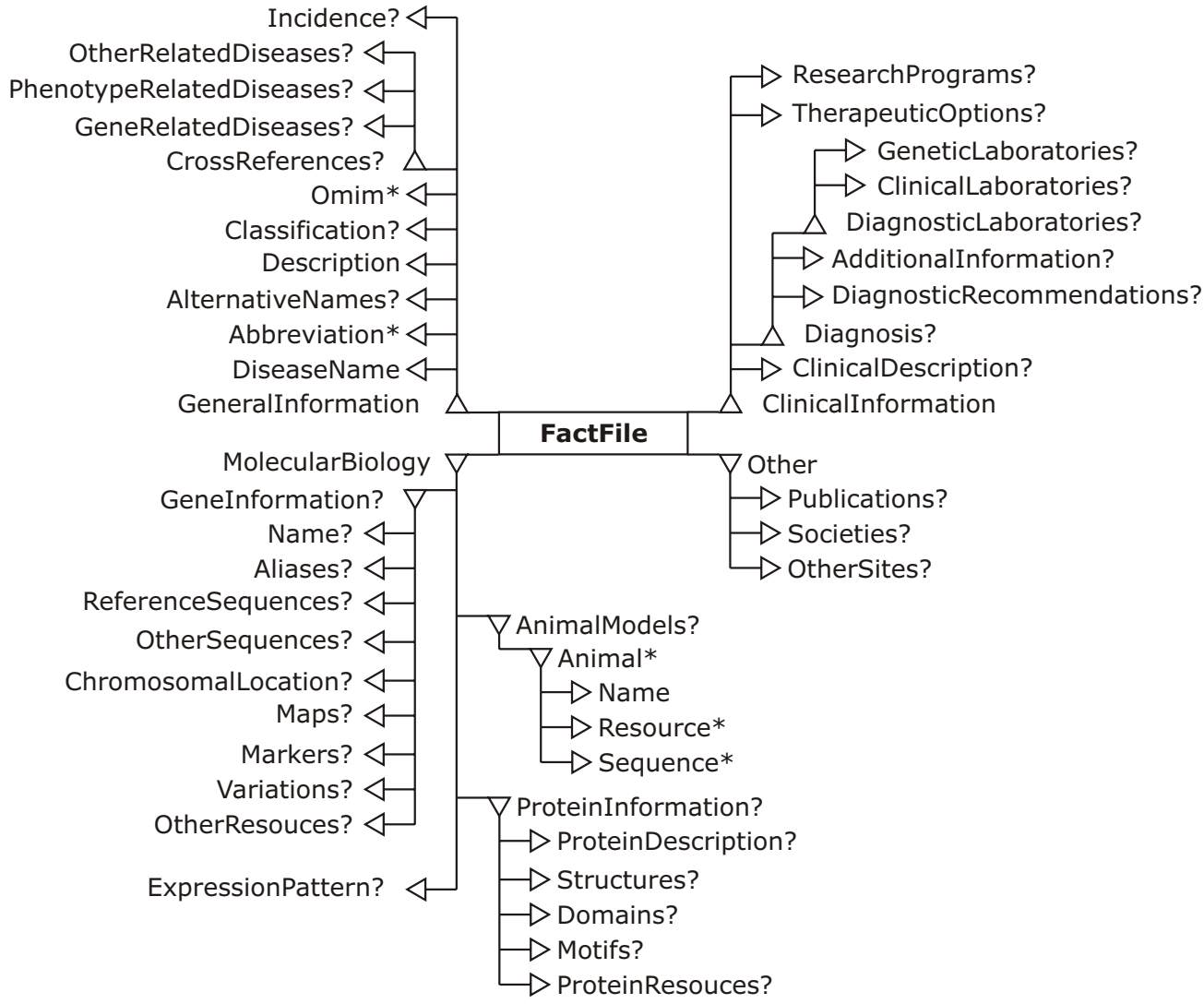

Figure 1
The tree diagram of main IDML elements of fact file data model. The operators "?" (optional), "*" (zero or more) and "+" (one or more) are used to denote cardinality indicating how many instances of an element type are permitted.

The fact file data model can be depicted as a tree structure graph where a <*FactFile*> element is a root (Figure 1). Fact files make use of the following specifications, standards and databases: HUGO nomenclature [25], RefSeq [22], Swiss-Prot [26] and SOURCE [27].

Stand-alone IDML fact files have been generated for each PID. The fact files are uniquely identified by an *id* attribute of *FactFile* root element. The major concepts in the first tier of the fact file hierarchy below the root level are general information, clinical information and molecular biology (Table 1). In addition, there are other resources, which contain links to related information providers. Each of these elements, in turn, comprises one or more additional levels of guideline constructs.

The components of the fact file model are defined as IDML elements. According to XML, elements have distinct names and they are delimited with start and end tag, e.g. <*DiseaseName*>X-linked agammaglobulinemia</*DiseaseName*>. Elements may contain other elements or attributes, they may store text, or they may be empty.

Table 1: The description of high-level concepts in the fact file document model

Element	Description	Content	Content model[a]
FactFile	The root element for IDML-based fact file document	Elements	(GeneralInformation, ClinicalInformation, MolecularBiology, Other)
GeneralInformation	Describes the disease in general terms	Elements	(DiseaseName, Abbreviation*, AlternativeNames?, Description, Classification?, Omim*, CrossReferences?, Incidence?)
ClinicalInformation	The short overview of characteristic clinical features	Elements	(ClinicalDescription?, Diagnosis?, TherapeuticOptions?, ResearchPrograms?)
MolecularBiology	Molecular genetic elements	Elements	(GeneInformation?, AnimalModels?, ProteinInformation?, ExpressionPattern?)
Other	Other related information	Elements	(Publications?, Societies?, OtherSites?)

[a] Content model consists of a set of parenthesis enclosing some combination of child element names and operators. The order operators "," (strict sequence) and "|" (choice) indicate how elements may be combined. The operators "?" (optional), "*" (zero or more) and "+" (one or more) are used to denote cardinality indicating how many instances of an element type are permitted?

Table 2: The description of IDML: GeneralInformation element

Element	Description	Content	Content model[a]
DiseaseName	Disease name	Type	String
Abbreviation	Abbreviation for disease name	Type	String
AlternativeNames	List of alternatively used disease names	Elements	(Name*)
Description	General description of disease	Mixed	(Glink | Italic)*
Classification	Classifies document explicitely in the fact files hierarchy	Elements	(Class)
Omim	A collection of the related references to the OMIM database	Elements	(OmimReference+)
CrossReferences	Refers to the related fact files	Elements	(PhenotypeRelatedDiseases?, OtherRelatedDiseases?, GeneRelatedDiseases?)
Incidence	Description of incidence	Type	String

[a] See table 1

Elements may appear as often as required. Many IDML elements contain *href* attribute for hyperlinking to more detailed information by using globally unique idenfier URL (Unified Resource Locator). The element naming convention follows the approach used by Electronic Business XML (ebXML) core components [28]. The IDML specified element names are in upper camel case (*UpperCamelCase*) and attribute names are in lower camel case (*lowerCamelCase*) notations. The usage of acronyms has been avoided, but when they are used the capitalization remains (example: *ReferenceDNA*).

General information elements identify a particular genetic disease and describe pattern of heritance and frequency in general terms (Table 2). The <*Abbreviation*> element includes commonly used abbreviations and the <*AlternativeNames*> element lists known aliases and synonymous names for the disease. The <*Description*> element provides a short overview in general terms. The description text may contain several *Glink* -tagged words that can act as links to a glossary, which is an integral part of the IDR service. The <C *lassification*> element is used to classify a disease explicitly to a group of related diseases. It exploits the hierarchic structure of XML documents by nesting <*Class*> elements. Each <*Class*> element contains a unique identifier in *level* attribute and a class name in <*Title*> element. The <*Omim*> element links the fact file to the Online Mendelian Inheritance in Man (OMIM) knowledge base [4] and the <*CrossReferences*> element refers to the related fact files grouped in <*Phenotype*>, <*Gene*>, and <*OtherRelatedDiseases*> elements. The incidence element stores information about disease frequency in human populations.

Table 3: The description of IDML: ClinicalInformation element

Element	Description	Content	Content model[a]
ClinicalDescription	Describes characteristic clinical features	Mixed	(Glink | Italic)*
Diagnosis	A collection of diagnostic guidelines and laboratories	Elements	(DiagnosticRecommendations?, AdditionalInformation?, DiagnosticLaboratories?)
TherapeuticOptions	A collection of available therapeutic options	Elements	(Option+)
ResearchPrograms	A collection of related studies		(Program+)

[a] See table 1

Table 4: The description of IDML: MolecularBiology element

Element	Description	Content	Content model[a]
GeneInformation	Contains information on the gene name, aliases, reference sequences, chromosomal location, maps, markers, variations and other gene related resources	Elements	(Name?, Aliases?, ReferenceSequences?, OtherSequences?, ChromosomalLocation?, Maps?, Markers?, Variations?, OtherResources?)
AnimalModels	A collection of related transgenic animal studies	Elements	(Animal*)
ProteinInformation	Contains information on protein characteristic features, structures, domains, motifs and other protein resources	Elements	(ProteinDescription?, Structures?, Domains?, Motifs?, ProteinResources?)
ExpressionPattern	Gene expression levels in a variety of cells and tissues	Elements	(Expression*)

[a] See table 1

Clinical information elements provide a short overview of characteristic clinical features, diagnosis, treatment and research related to the disease (Table 3). The <*ClinicalDescription*> element stores text, that describes characteristic clinical features and the most important laboratory findings. The <*Diagnosis*> refers to data on diagnostic criteria and guidelines. It also refers to databases of laboratories performing clinical and/or genetic analyses for the disease including IDdiagnostics [29], the European Directory of DNA diagnostic Laboratories (EDDNAL, http://www.eddnal.com) and GeneTests [30]. Detailed diagnostic guidelines are available for several IDs [31]. The <*TherapeuticOptions*> lists therapeutic interventions that are available. The <*ResearchPrograms*> includes important research and clinical trials related to the disease.

Molecular biology comprises the main genetic components on DNA, RNA and protein level, animal models, protein properties and expression patterns (Table 4). The <*GeneInformation*> elements store the basic information on gene names, aliases and synonyms. They provide also <*ReferenceSequences*> element that covers reference sequences on three levels and lists also other related sequences that are available from sequence databanks. Information on gene locus is stored in <*ChromosomalLocation*>, <*Maps*>, and <*Markers*> elements. The <*Variations*> element refers to related locus specific mutation and single nucleotide polymorphism (SNP) databases. We and others are maintaining a large number of immunodeficiency mutation databases [24,32]. The <*OtherResources*> element refers to the other genetic web services such as Ensembl [33], GENATLAS [34], GeneCards [20], UniGene [21], LocusLink [22], euGenes [35], GDB [36], GeneLynx [37] and SOURCE [27]. The <*AnimalModels*> element refers to the related transgenic animal studies.

The <*ProteinInformation*> element stores characteristic structural and functional properties of the protein. The <*ProteinDescription*> contains several subelements e.g. <*Function*>, <*SubcellularLocation*>, <*CatalyticActivity*>, which are inherited from the Swiss-Prot entry model [26]. The <*Structures*> element refers to solved protein structures available in Protein DataBank (PDB) [38]. The domain and motif elements describe conserved protein regions. Each <*Domain*>, <*Motif*> and further <*ProteinResources*> element includes links to related resources for example in Pfam [39], InterPro [40], ProDom [41], SMART [42] or PROSITE [43]. The <*ExpressionPattern*> stores information on gene or protein expression. This information is mainly from SOURCE [27], which is a

Table 5: The description of IDML: Other element

Element	Description	Content	Content model[a]
Publications	A collection of related publications	Elements	(PubmedSearch?, Pubmed?)
Societies	List of related general and disease specific societies	Elements	(GeneralSocieties?, DiseaseSpecificSocieties?)
OtherSites	A collection of other related Web sites	Elements	(Site+)

[a] See table 1

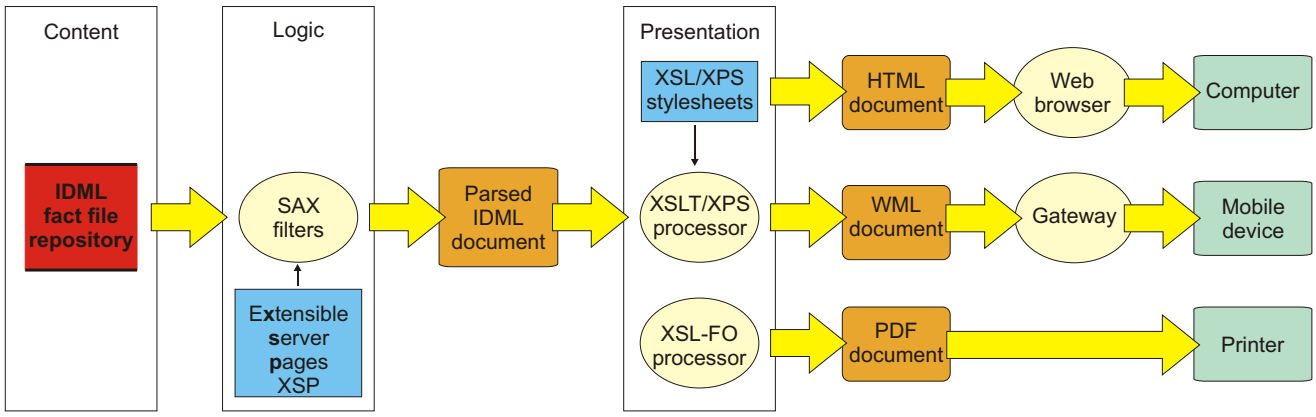

Figure 2
The data flow diagram of IDR. Notations: A red open-ended rectangle represents data store, light yellow ovals represent processes, blue rectangles represent external entities, orange rounded rectangles shows transferred documents, green rounded rectangles represent destination devices and arrows shows the flow of information from its source to its destination. White rectangles show the separation of content, logic and presentation.

web-based resource bringing together genetic information from different sources.

The last high level element <*Other*> stores various information in elements such as <*Publications*> and <*Societies*>, which is categorized by <*GeneralSocieties*> and <*DiseaseSpecificSocieties*> elements (Table 5). The <*OtherSites*> element refers to other related resources in the Internet.

The IDML schema version 1.0 (idml.xsd file), examples of IDML-document and documentation on the syntax are available at our web site http://bioinf.uta.fi/idml/. The IDML document type definition file (idml.dtd) is also available, althougth we prefer to use the IDML schema for validation. Many IDML elements are optional. The syntax allows one to put comments, both within and outside of the XML markup. The parser must pass internal comments to the application programs, which can then properly treat the information. IDML documents specify which version of the schema is to be used to validate their content, eliminating possible confusion when several versions exist. IDML is open access, however, a licence is needed for building other services. Contact the authors for details.

Results

The IDML model was implemented to describe primary immunodeficiencies, which is a group of over 100 hereditary diseases. IDs can be grouped as follows: combined B and T cell immunodeficiencies, deficiencies predominantly affecting antibody production, defects in lymphocyte apoptosis, other well-defined immunodeficiency syndromes, defects of phagocyte function, interferon-γ (IFNγ) associated immunodeficiencies, DNA breakage associated syndromes, defects of the complement cascade proteins, and defects of complement regulatory proteins. The disease information is stored in IDML-based fact files, which form the central repository for data retrieval of ImmunoDeficiency Resource (IDR) service [23,24,29] available at http://bioinf.uta.fi/idr/. The data flow diagram of IDR is shown in figure 2. In addition to information on fact files, the IDR contains several introductory texts and collections of immunology related data sources.

The IDR pages are extensively hyperlinked to our on-line immunology glossary. More detailed description about the IDR web service has been published elsewhere [23,24].

The ImmunoDeficiency Resource is a comprehensive knowledge base on immunodeficiencies. IDR is developed and maintained by IMT Bioinformatics group in collaboration with experts on individual immunodeficiencies. All the information in the IDR will be validated by expert curators. However, all changes, additions and corrections to the fact files are made by our group. IDR is designed for different user groups such as researchers, physicians and nurses as well as patients and their families and the general public. IDR contains fact files for practically all known PIDs. The numerous individual data items in IDR have been collected partly manually, usually with simple Perl scripts written for datamining from numerous local and Internet databases and services.

We selected Apache AxKit XML Application Server version 1.61 for implementation of the IDML-encoded web service. AxKit is an application and document server that uses XML processing pipelines to generate and process content and to deliver it to clients in a wide variety of formats, such as HTML, WML, PDF and plain text using either standard techniques of World Wide Web Consortium (XSLT) [13], or flexible custom codes (XPathScript XPS, eXtensible Server Pages XSP).

Similar XML application server called Cocoon, has been written in Java. We settled on AxKit, because it is built in Perl, which makes it easy to integrate with bioinformatic applications many of which are written in Perl. It is important to note that AxKit is not limited to XML source documents. Non-XML documents and data sources can be converted to XML when necessary. AxKit separates the content, logic and presentation. The content reuse was implemented with XInclude [44] and XPointer [45] techniques. The root element of IDML schema is <*FactFiles*> and according to W3C Recommendation "Namespaces in XML" [46] we declared a default namespace attribute in the root element xmlns:idml="http://bioinf.uta.fi/idml" to avoid the problems of ambiquity and name collisions.

Each fact file is stored in an IDML file, that has a unique name and url address. When a fact file requests the pipeline it might look like this in diagramatic terms : Request > [XSP] > (XML) > [XSLT] > (HTML) > Browser, where processors are in square brackets and products in round brackets. The output of XSP pages is structured XML content, which can pipe through XSLT to produce HTML. The XSP feature is not currently in use in the IDR.

The information on fact files can be easily transformed and presented in any platform. It is easy to write platform or even browser and screen specific pages. We have implemented a transformation from IDML to WML for portable devices (such as mobile phones) with WAP compliance (Wireless Application Protocol). The fact files are available via bioinformatics related WAP service, BioWAP [47,48]http:///bioinf.uta.fi/wml/welcome.wml practically anywhere, anytime.

New web techniques are developed continuously. During this project a number of new specifications and software appeared, requiring upgrading of the system many times. The separation of content and presentation enables to share the project for people who are responsible for information content and people who develop the knowledge management techniques. Once the data model was created, we have not had to touch it hardly at all in spite of technical improvements, content additions and deletions.

Discussion

As far as we know there are no other efforts to develop a markup language to describe connections between disease and genetic information. The IDML was designed with following purposes in mind. First, we wanted a markup that is able to present disease, clinical, diagnostic and genetic information and relations between them. Secondly, the data model structure had to be intuitive, hierarchical, flexible, but still machine and human readable. Sometimes the relatively large XML files can appear verbose for human readers, but hierarchically and logically organized structure in addition to semantic markup facilitate the interpretation of documents. Thirdly, an application and platform independent data format was needed. Its portability, extensibility and robustness are primary advantages for interoperating heterogeneous systems. The availability of open source and free tools for processing files in all major programming languages is important. The openness of source code as well as data formats and data itself allows better integration and interoperation between data resources. The IDML enables the seamless integration of genetic and disease information resources in the Internet. The data model is appropriate for the implementation of automated decision support systems such as diagnostic consultations. Fourthly, the data have to be unambiguous and validated.

A fact file is a user oriented user interface, which serves as a good starting point to explore information on hereditary diseases. For some time now, there has been many advanced search facilities in the Internet such as Google, that are able to find very fast web pages that contain given keywords. However, the web searches typically turn up innumerable completetely irrelevant "hits", requiring

much manual filtering by the user. Navarro *et al*. lists some issues related to database searching and accessibility that can cause difficulties [49] including inaccurate and redundant search results, nomenclature issues, lack of internal access, non-availability of the source code, lack of customization and differing data formats. New methods are needed for improving search results.

There is an increasing number of biomedical data sources in the Internet. The Human Genome Initiative [2] and other genome research projects have generated enormous quantities of data. The genetic data is well organized in web accessible databases for example EMBL [50], GenBank [51], Swiss-Prot [26], etc. Several organizations offer public interfaces for obtaining biomedical information across a range of domains. They provide numerous tools and applications for genetic data retrieval and analysis for example with Sequence Retrieval System SRS [52] and BioPerl [53]. In addition to the sequence information, databases contain a lot of valuable information in annotations. There are also some genetic knowledge bases such as GeneCards and GeneLynx that comprise the essential information on genes. Swiss-Prot contains also some disease related annotations. The most comprehensive database on hereditary diseases is OMIM [4], which contains descriptions for known hereditary diseases.

Almost all pages in the Internet have been written in HyperText Markup Language (HTML) where it is used for style description. It provides some possibilities for simple description about a document. It is able to use special metatags that contain simple keywords or more advanced descriptions like Dublin Core Languages, but they are very little utilised and only the most sophisticated search engines can exploit them.

There are some efforts to integrate heterogeneous biomedical databases [15,54,55]. Some level of standardization is required for more automatic integration. Development of integration techniques is moving databases towards the Internet and XML-based systems [56]. In the future, Web services will use standard Internet protocols including SOAP, WSDL, and UDDI for interoperability with other resources. Thereby the flexible and expandable integration of diverse scientific tools will be achieved.

Conclusion

The XML-based language IDML and fact file data model were developed for integrating, storing and exchanging information on inherited diseases. The IDML language and fact file model are implemented in the IDR knowledge base. The fact files can be easily transformed from IDML to any format such as HTML or WML using either standard W3C techniques or flexible custom code. The content management as well as the exchange of presentation are facilitated by separating document content and presentation. The IDML-based information system was proved to be a viable and applicable specification for inherited diseases. Numerous downloads (altogether more than 250,000) from the IDR knowledge base during the last two years have proved the applicability and adaptability of the fact file model.

List of abbreviations

AxKit An XML Delivery Toolkit for Apache

BioWAP Bioinformatics service for portable devices

DOM Document Object Model

DTD Document Type Definition

ebXML Electronic Business XML

EDDNAL European Directory of DNA Diagnostic Laboratories

EMBL Genetic sequence database by European Molecular Biology Laboratory

GDB Genome Database

GenBank Genetic sequence database by National Center for Biotechnology Information

HTML Hypertext Markup Language

IDML Inherited Disease Markup Language

IDR ImmunoDeficiency Resource

IE Information extraction

IFNγ Interferon gamma

NLP Natural language processing

OMIM Online Mendelian Inheritance in Man

PID Primary Immunodeficiency

PDB Protein DataBank

PDF Portable Document Format

Pfam Protein Families Database

ProDom Protein Domain Database

PROSITE Database of Protein Families and Domains

RDF Resource Description Framework

SGML Standard Generalized Markup Language

SMART Simple Modular Architecture Research Tool

SNP Single nucleotide polymorphism

SOAP Simple Object Access Protocol

SOURCE Genomic resource in the Internet

Swiss-Prot Protein knowledgebase

UDDI Universal Description, Discovery, and Integration

URL Unified Resource Locator

W3C World Wide Web Consortium

WAP Wireless Application Protocol

WML Wireless Markup Language

WSDL Web Services Definition/Description Language

XML Extensible Markup Language

XPS XPathScript

XSL Extensible Style Language

XSLT XSL Transformations

XSP eXtensible Server Pages

Competing interests

The author(s) declare that they have no competing interests.

Additional material

Additional File 1

An XML schema for IDML
Click here for file
[http://www.biomedcentral.com/content/supplementary/1472-6947-5-21-S1.xsd]

Additional File 2

A DTD file for IDML
Click here for file
[http://www.biomedcentral.com/content/supplementary/1472-6947-5-21-S2.dtd]

Additional File 3

An example file using IDML. A fact file for X-linked agammaglobulinemia.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1472-6947-5-21-S3.xml]

Acknowledgements

Financial support from the European Union, the National Technology Agency of Finland and the Medical Research Fund of Tampere University Hospital is gratefully acknowledged.

References

1. Collins FS, Green ED, Guttmacher AE, Guyer MS: **A vision for the future of genomics research.** *Nature* 2003, **422:**835-847.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, Szustakowki J, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
3. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill

M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.

4. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2002, **30:**52-55.
5. Guttmacher AE, Collins FS: **Genomic medicine - a primer.** *N Engl J Med* 2002, **347:**1512-1520.
6. Maojo V, Kulikowski CA: **Bioinformatics and medical informatics: collaborations on the road to genomic medicine?** *J Am Med Inform Assoc* 2003, **10:**515-522.
7. Grivell L: **Mining the bibliome: searching for a needle in a haystack? New computing tools are needed to effectively scan the growing amount of scientific literature for useful information.** *EMBO Rep* 2002, **3:**200-203.
8. Yandell MD, Majoros WH: **Genomics and natural language processing.** *Nat Rev Genet* 2002, **3:**601-610.
9. World Wide Web Consortium: **Extensible Markup Language (XML) 1.0 (Second Edition).** [http://www.w3.org/TR/REC-xml].
10. International Organization for Standardization: **Standard Generalized Markup Language (SGML).** [http://www.iso.org/iso/eCatogueDetailPage.CatalogueDetail?CSNUMBER=16387&ICS1=35&ICS2=240&ICS3=30].
11. World Wide Web Consortium: **Document Object Model (DOM) Level 1 Specification.** [http://www.w3.org/TR/REC-DOM-Level-1/].
12. World Wide Web Consortium: **XML Schema Part 0: Primer.** [http://www.w3.org/TR/xmlschema-0/].
13. World Wide Web Consortium: **XSL Transformations (XSLT) Version 1.0.** [http://www.w3.org/TR/xslt].
14. World Wide Web Consortium: **RDF/XML Syntax Specification (Revised).** [http://www.w3.org/TR/rdf-syntax-grammar/].
15. Mork P, Halevy A, Tarczy-Hornoch P: **A model for data integration systems of biomedical data applied to online genetic databases.** *Proc AMIA Symp* 2001:473-477.
16. Achard F, Vaysseix G, Barillot E: **XML, bioinformatics and data integration.** *Bioinformatics* 2001, **17:**115-125.
17. Coyle JF, Mori AR, Huff SM: **Standards for detailed clinical models as the basis for medical data exchange and decision support.** *Int J Med Inf* 2003, **69:**157-174.
18. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, Forrey A, Mercer K, DeMoor G, Hook J, Williams W, Case J, Maloney P: **LOINC, a universal standard for identifying laboratory observations: a 5-year update.** *Clin Chem* 2003, **49:**624-633.
19. Lee KP, Hu J: **XML Schema Representation of DICOM Structured Reporting.** *J Am Med Inform Assoc* 2003, **10:**213-223.
20. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D: **GeneCards: encyclopedia for genes, proteins and diseases.** [http://bioinformatics.weizmann.ac.il/cards].
21. Schuler GD: **Pieces of the puzzle: expressed sequence tags and the catalog of human genes.** *J Mol Med* 1997, **75:**694-698.
22. Pruitt KD, Maglott DR: **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 2001, **29:**137-140.
23. Väliaho J, Pusa M, Ylinen T, Vihinen M: **IDR: the ImmunoDeficiency Resource.** *Nucleic Acids Res* 2002, **30:**232-234.
24. Väliaho J, Riikonen P, Vihinen M: **Novel immunodeficiency data servers.** *Immunol Rev* 2000, **178:**177-185.
25. Wain HM, Lush MJ, Ducluzeau F, Khodiyar VK, Povey S: **Genew: the Human Gene Nomenclature Database, 2004 updates.** *Nucleic Acids Res* 2004, **Database issue:**D255-7.
26. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31:**365-370.
27. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO, Alizadeh AA: **SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data.** *Nucleic Acids Res* 2003, **31:**219-223.
28. Eisenberg B, Nicull D: **ebXML Technical Architecture Specification v1.0.4.** [http://www.ebxml.org/specs/ebTA.pdf].
29. Samarghitean C, Valiaho J, Vihinen M: **Online Registry of Genetic and Clinical Immunodeficiency Diagnostic Laboratories, IDdiagnostics.** *J Clin Immunol* 2004, **24:**53-61.
30. Tarczy-Hornoch P, Shannon P, Baskin P, Espeseth M, Pagon RA: **GeneClinics: a hybrid text/data electronic publishing model using XML applied to clinical genetic testing.** *J Am Med Inform Assoc* 2000, **7:**267-276.
31. Conley ME, Notarangelo LD, Etzioni A: **Diagnostic criteria for primary immunodeficiencies.** *Clin Immunol* 1999, **93:**190-197.
32. Vihinen M, Arredondo-Vega FX, Casanova JL, Etzioni A, Giliani S, Hammarström L, Hershfield MS, Heyworth PG, Hsu AP, Lähdesmäki A, Lappalainen I, Notarangelo LD, Puck JM, Reith W, Roos D, Schumacher RF, Schwarz K, Vezzoni P, Villa A, Väliaho J, Smith CI: **Primary immunodeficiency mutation databases.** *Adv Genet* 2001, **43:**103-188.
33. Birney E, Andrews D, Bevan P, Caccamo M, Cameron G, Chen Y, Clarke L, Coates G, Cox T, Cuff J, Curwen V, Cutts T, Down T, Durbin R, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz H, Iyer V, Kahari A, Jekosch K, Kasprzyk A, Keefe D, Keenan S, Lehväslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark C, Clamp M, Hubbard T: **Ensembl 2004.** *Nucleic Acids Res* 2004, **Database issue:**D468-70.
34. Frezal J: **Genatlas database, genes and development defects.** *C R Acad Sci III* 1998, **321:**805-817.
35. Gilbert DG: **euGenes: a eukaryote genome information system.** *Nucleic Acids Res* 2002, **30:**145-148.
36. **GDB Human Genome Database** [http://www.gdb.org/]
37. Lenhard B, Hayes WS, Wasserman WW: **GeneLynx: a gene-centric portal to the human genome.** *Genome Res* 2001, **11:**2151-2157.
38. Bernstein FC, Koetzle TF, Williams GJ, Meyer EFJ, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: **The Protein Data Bank: a computer-based archival file for macromolecular structures.** *J Mol Biol* 1977, **112:**535-542.
39. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2002, **30:**276-280.
40. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJ, Vaughan R,

Zdobnov EM: **The InterPro Database, 2003 brings increased coverage and new features.** *Nucleic Acids Res* 2003, **31:**315-318.
41. Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D: **ProDom: automated clustering of homologous domains.** *Brief Bioinform* 2002, **3:**246-251.
42. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P: **SMART 4.0: towards genomic data integration.** *Nucleic Acids Res* 2004, **Database issue:**D142-4.
43. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P: **PROSITE: a documented database using patterns and profiles as motif descriptors.** *Brief Bioinform* 2002, **3:**265-274.
44. World Wide Web Consortium: **XML Inclusions (XInclude) Version 1.0.** [http://www.w3.org/TR/xinclude/].
45. World Wide Web Consortium: **XML Pointer Language (XPointer).** [http://www.w3.org/TR/xptr/].
46. World Wide Web Consortium: **Namespaces in XML.** [http://www.w3.org/TR/REC-xml-names/].
47. Riikonen P, Boberg J, Salakoski T, Vihinen M: **BioWAP, mobile Internet service for bioinformatics.** *Bioinformatics* 2001, **17:**855-856.
48. Riikonen P, Boberg J, Salakoski T, Vihinen M: **Mobile access to biological databases on the Internet.** *IEEE Trans Biomed Eng* 2002, **49:**1477-1479.
49. Navarro JD, Niranjan V, Peri S, K. JC: **From biological databases to platforms for biomedical discovery.** *Trends in Biotechnol* 2003, **21:**263-268.
50. Kulikova T, Aldebert P, Althorpe N, Baker W, Bates K, Browne P, van den Broek A, Cochrane G, Duggan K, Eberhardt R, Faruque N, Garcia-Pastor M, Harte N, Kanz C, Leinonen R, Lin Q, Lombard V, Lopez R, Mancuso R, McHale M, Nardone F, Silventoinen V, Stoehr P, Stoesser G, Tuli MA, Tzouvara K, Vaughan R, Wu D, Zhu W, Apweiler R: **The EMBL Nucleotide Sequence Database.** *Nucleic Acids Res* 2004, **Database issue:**D27-30.
51. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank: update.** *Nucleic Acids Res* 2004, **Database issue:**D23-6.
52. Zdobnov EM, Lopez R, Apweiler R, Etzold T: **The EBI SRS server-new features.** *Bioinformatics* 2002, **18:**1149-1150.
53. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Res* 2002, **12:**1611-1618.
54. Lawrence R, Barker K: **Integrating relational database schemas using a standardized dictionary: ; Las Vegas, Nevada, United States.** ACM Press; 2001:225-230.
55. Stevens RD, Robinson AJ, Goble CA: **myGrid: personalised bioinformatics on the information grid.** *Bioinformatics* 2003, **Suppl 1:**I302-I304.
56. Das M, Lawhead PB: **Information storage and management in large web-based applications using XML.** *J Comput Small Coll* 2003, **18:**72-79.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1472-6947/5/21/prepub