Designing Usable Bio-information Architectures

Davide Bolchini[1], Anthony Finkestein[2], and Paolo Paolini[3]

[1] Indiana University, School of Informatics (IUPUI)
dbolchin@iupui.edu
[2] University College London, Dept. of Computer Science
a.finkeslstein@cs.ucl.ac.uk
[3] Politecnico di Milano, Dept. of Information and Electronics
paolo.paolini@polimi.it

Abstract. Bioinformatics websites offer to the life science large community repositories of information ranging from genes, genomes, proteins, experimental data and their integration, with the aim of supporting the elucidation of biological processes. As the bioinformatics community increasingly relies on the design, sharing and use of web-based resources, it is important to systematically address the usability of these applications and to deliver a more rewarding user experience to researchers. The bioinformatics community is also acknowledging the role that Human-Computer Interaction can play to improve the usability of these systems. In the context of a project aiming at improving the usability of large bioinformatics websites, we carried out an in-depth usability analysis and conceptual redesign of a well-known protein repository, with the aim of characterizing information architecture usability problems and providing corresponding design solutions to improve the user experience. This design has been validated and refined using interactive prototypes with users, usability experts and domain experts, and opens a new set of navigation opportunities which has the potential to improve the research work of bioinformaticians. Although being a preliminary study, the research reveals generic information architecture and navigation issues which have design implications for browsing-intensive bioinformatics repositories at large.

Keywords: usability, information architecture, navigation design, bioinformatics.

1 Introduction

Bioinformatics web applications are developed to offer to the life science research community up-to-date repositories containing information ranging from genes, genomes, proteins, experimental data and their integration, with the aim of supporting the elucidation of biological processes. As the bioinformatics community increasingly relies on the design, sharing and use of web-based resources, it is important to systematically address the usability of these applications and to deliver a more rewarding user experience to researchers [5]. The bioinformatics community is also increasingly acknowledging the importance of the role that Human-Computer Interaction can play to improve the utility and usability of these systems [1][4].

A specific family of web repositories is devoted to offer a collection of all known proteins, and classifying them according to their structure. Within a wider project

J.A. Jacko (Ed.): Human-Computer Interaction, Part IV, HCII 2009, LNCS 5613, pp. 653–662, 2009.

aimed at understanding and characterizing general usability issues of web-based bioinformatics resources and provide design improvements, we have carried out an in-depth usability analysis and conceptual redesign of a well-known protein classification repository (CATH, http://www.cathdb.info/). Through a detailed usability inspection [3][6] of the information architecture currently supported, we have discovered that, being based on a hierarchical classification of the content, the navigation architecture follows a strict tree-based hierarchical paradigm, where each branch of the tree determines a sub-class of the previous branch, and where each element in the hierarchy has one and only one position in the tree. This hierarchical classification of the proteins is used as the main navigation paradigm to access and explore the protein collection. This situation has a potential negative impact on the overall usability of the application, as it hinders the possibility of exploring and navigating the classification and protein collection with efficiency and flexibility, and poses obstacles to the accomplishments of simple, basic exploratory user tasks (e.g. the possibility to browse proteins without being forced to specify beforehand 8 different parameters).

To address these issues and propose an enhanced design solution which can be useful and applicable to a variety of bioinformatics resources using with similar design solutions, we have proposed a fundamental paradigm shift in the design of the browsing experience, by capitalizing on well-known design principles from the hypermedia and web communities. The basic navigation paradigm that we have proposed for CATH is based on the assumption that each classification criterion for the proteins (e.g. class, topology, architecture, homologous superfamily) can be modelled as a primary navigation dimension (facet or trail), to be browsed orthogonally to all the others, instead of being represented just as a level of the hierarchy. Using each classification criterion as an independent navigation driver, it is possible to make these criteria interact, enabling the users to visualize, explore and browse the relationships between them with a greater flexibility than the one currently offered by a hierarchical navigation paradigm.

This design, which have been validated and iteratively refined on interactive prototypes with users, usability experts and domain experts (bioinformaticians), opens a whole set of new navigation possibilities which improve the quality of the overall user experience and which will be reported in this paper.

The remainder of the paper is organized as follows. Section 2 overviews the basic related work and underlying body of knowledge that supports the redesign work described in the paper, and mainly the design principles that have been used. Section 3 shows the key results of the preliminary usability inspection of CATH; which identified a number of usability and the need for a conceptual redesign. Section 4 illustrates the conceptual remodeling of the CATH user experience, which led to the development of specific design proposals, briefly illustrated and discussed in Section 5. Section 6 summarizes the contribution of the work and points to relevant research outlooks.

2 Theoretical Background and Related Work

There is evidence of an increasing awareness of the need of usability studies in the development of biomedical systems in general and of the benefits that a systematic user-centered design process can bring to the development of interactive systems in the

bio-related areas. There have also been some initial but notable efforts to address the usability of bioinformatics web-based applications [6]. The challenge of bringing an increased awareness of usability and user-centered design to the development of bioinformatics applications is currently tackled from different disciplinary perspectives.

The Human-Centered Software Engineering (HCSE) at Concordia University has worked on developing integrated web-based interfaces to popular bioinformatics portals in order to provide integrated access to web resources relevant to a set of typical tasks [5]. Acknowledging the fact that web bioinformatics resources are so diversified and scattered around - thus forcing researchers to discover, locate these resources and then learn different interaction paradigms to access, search for data and complete tasks - this research explored the possibility of offering a unique one-stop access interface to a selected (limited) set of recurrent bioinformatics tasks.

Approaching the complexity of bioinformatics resources from another perspective, the Human-Computer Interaction Lab at the University of Maryland is investigating advanced visualization techniques to access and manipulate large multimedia information sets in biological databases [4]. Usability challenges for complex data visualization and exploration in support to discovery and decision-making in bioinformatics are tackled with advanced visualization paradigms. Moving to a higher level of user activities, Joan Bartlett at McGill University has been investigating the daily activities of bioinformatics researchers in order to derive a list of typical information tasks that entail the use of web-based resources to complete [1].

Although these contributions cover important aspects of improving the user experience of biological databases little has been done to analyse the underlying *design characteristics* of web bioinformatics resources that can lead to potential usability problems. Tackling design issues captures the usability problems at their source, thus providing strategies to prevent the emergence of problems in current and future applications. Recently, a contribution in this direction has been elaborated by the authors in their preliminary work in collecting and characterizing general types of usability problems in web-based bioinformatics repositories [6]. Given the centrality of the attention to the quality of design as one important factor determining the quality of the user experience, the underlying theoretical and methodological foundation for this recent advance in the field, and also for the research presented in this paper, lies in the long-standing tradition of conceptual hypermedia and web modeling. Over the last decade, a rich body of knowledge aimed at providing the conceptual tools to design, analyze and describe the structure of complex hypertext and hypermedia applications (and consequently the ones available on the web) has been produced in hypertext, hypermedia and web engineering communities. A constant line of research that can be identified is the one related to design models, i.e. systematic and cohesive set of modeling abstractions useful to describe the complexity of an interactive application at the proper level of granularity, with the goal of making this complexity more understandable and tractable for various purposes (mainly design, analysis and evaluation).

One of the key underpinning theoretical construct of these models is the distinction between the design of the content or information base (also called “hyperbase”) and the design of the access structures (navigational paths enabling the users to locate and reach the content of interest). Recently, one of the lasts heir of this research tradition is IDM (Interactive Dialogue Model), which offers a basic, lightweight set of modeling primitives to design the information and navigation patterns of a complex

interactive applications in terms of dialogue between the user and the system [2]. A common denominators of this rich (and growing) body of knowledge in the areas of conceptual design, design primitives, and modeling notations for complex interactive applications, is the assumption that the design of the user experience involves a set of subjective (although reasoned an grounded) design decisions driven by the application goals and the user needs and tasks, and not by a predefined organization of the information available in the knowledge domain. A quite recent and related line of work, ultimately stemming from the information architecture and library science tradition, but conceptually very similar to the early advancement of hypertext and hypermedia design models, is the definition of flexible navigation paradigms based on faceted classification [7], one of whose tenets is that multiple access paths can be designed to reach the same information object, and this provides more flexibility in effectively supporting user tasks.

This fundamental principle led to the definition of methods and tools to design navigation design patterns, access structures to content and information architectures that might support a flexible set of user tasks and scenarios, and are not primarily driven by an *a priori* information structure defined independently from the user experience requirements. Given our recent research results, we claim that the adoption of these (and other) fundamental design principles used in the usability, hypermedia and web design research community can be applied to the design of bioinformatics web applications and can bring an enhanced level of usability and, ultimately, research productivity in this growing domain of web-based resources.

3 Usability Issues in Rigid Hierarchical Navigation

In the context of a project aiming at improving the usability of large bioinformatics websites, we carried out an in-depth usability inspection and conceptual redesign of a well-known protein web-based repository (CATH, http://www.cathdb.info/), with the aim of characterizing the nature of the usability problems and providing corresponding design solutions to improve the user experience. CATH is the primary online resource for protein domain classification and is developed, maintained and hosted at University College London, Biochemistry Department. CATH classifies protein domains (currently ca. 93'000) into 8 levels, 4 related to the similarity concerning the structural characteristics (mainly the shape) of the proteins (levels are, for example, class, architecture, or topology), and 4 related to the similarities in the sequence of amino acids. In particular, CATH classifies protein domains according to the following hierarchical levels (adapted from http://www.cathdb.info/): *Class, C-level* is determined according to the secondary structure composition and packing within the structure. *Architecture, A-level* describes the overall shape of the domain structure as determined by the orientations of the secondary structures. *Topology (Fold family), T-level* are grouped into fold groups at this level depending on both the overall shape and connectivity of the secondary structures. *Homologous Superfamily, H-level* groups together protein domains which are thought to share a common ancestor. Additional 4 levels (S,O,L,I,D), grouping proteins according their similarity in sequence.

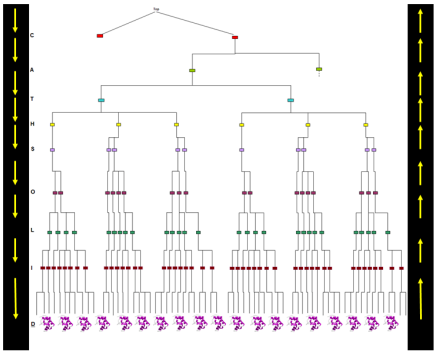

Fig. 1a. CATH hierarchical navigation

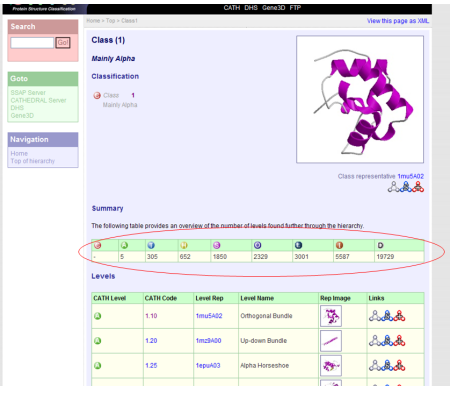

Fig. 1b. Interface over the hierarchy

This hierarchical classification of the protein domains is reflected by a tree-based hierarchical navigation structure as the only navigation paradigm offered by CATH to browse and explore the protein collection (Figure 1a). The information architecture designed is a classical 8-level hierarchy or tree, where each node has only one ancestor, i.e. it has a unique position in the hierarchy, and serves a mainly classification purpose, as the actual content (i.e. the detailed information about the protein domains) is contained in the leaf nodes (end points of the tree). We should note, however, that while the hierarchical classification of proteins based on their structure is *per se* a valuable knowledge asset for the life science community, the usability problems intrinsic to a large, multi-level tree-based navigation structure are several, and are here briefly summarized.

At each level of the hierarchy, access is granted to nodes to the immediate next level; whereas nodes further down on the hierarchy tree are not directly accessible (i.e., it is not possible to skip levels, see Figure 1b). This design poses severe obstacles to users who need to explore the protein classification by a specific criterion (Topology), from a given point in the tree, and are not interested in exploring intermediate classifications. In other words, this design forces users to traverse *all* the levels of the hierarchy to reach a protein domain of interest (leaf node). A necessary access sequence is imposed to the user's navigation within the hierarchy, which has the consequence of making the browsing flow very rigid, as potentially irrelevant steps are put on the way.

This navigation design might be effective *only* in scenarios in which users are able to clearly specify upfront the values of all (8) parameters of the hierarchy, in order to locate a protein domain. The solution, however, is some way far from being effective and efficient when users have more ill-defined knowledge of the classification parameters, need exploring and iteratively refining the browsing scope.

As additional usability problem, we should note that the rigidity imposed to the browsing mechanisms makes the user's interaction flow even more inefficient in case a sequence of branches in the hierarchy is minimally populated (one node for each level): this flattened classification makes the content base (leaf nodes) even more desirable for quick access but still users have to traverse all these one-node populated branches to reach the protein domains of interest.

4 Remodelling Hierarchies through Flexible Navigation Paradigms

To address these issues, we claim that the starting point to conceptualize the requirements for an enhanced design solution it to decouple two fundamental concerns during design: (a) *the Information Architecture paradigm*: a knowledge, domain-driven, purposeful representation of the information (e.g. hierarchical classification of the protein according to their structure), which can be useful *as is* for knowledge sharing, for scientific dissemination, or for very specific tasks; (b) *the Navigation/Interaction paradigms*, i.e. design strategies to support the user experience in terms of interaction and browsing possibilities on top of the existing information architecture. Many navigation styles can be supported on top of the same basic information architecture (including hierarchical ones). The design of the navigation paradigms should allow the users to browse and explore more efficiently and effectively the information architecture and the content offered, covering a wider range of potential tasks than the one supported by the constraints of the underlying information architecture.

The basic navigation design paradigm that we have proposed for CATH is based on the assumption that each classification criterion (class, topology, architecture, homologous superfamily), instead of being represented as a level of the hierarchy, can be modelled as a primary navigation dimension (facet or trail), to be browsed orthogonally to all the others.

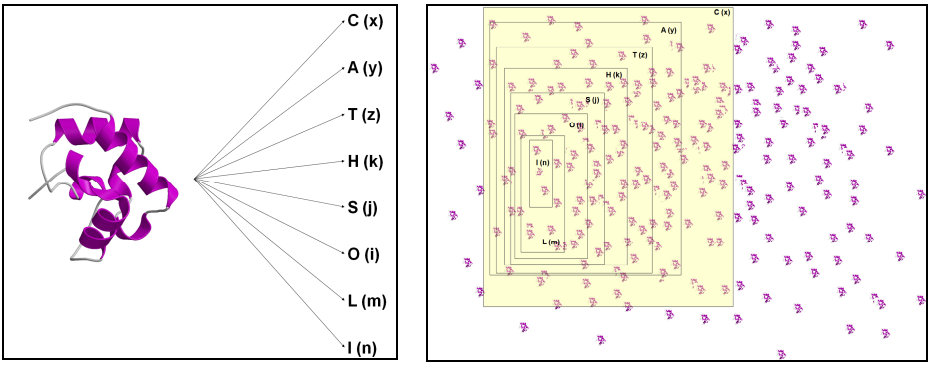

Fig. 2. Remodelling the Access Paths for the Protein Domains

Using each classification criterion as an independent navigation driver, it is possible to make these criteria interact, enabling the users to visualize, explore and browse the relationships between them with a greater flexibility than the one currently offered by a hierarchical navigation paradigm.

With this new modeling of the navigation space, the user can choose to browse the protein classification by any desired criteria (one of the 8 available in CATH), visualize the corresponding protein domain collection, and use the other criteria to refine or filter the browsing scope. The first result of the redesign is that, while the underlying information architecture remains strongly hierarchical, the navigation is remodeled as a semi-flat structure composed by the set of all classification criteria (called “facets”,

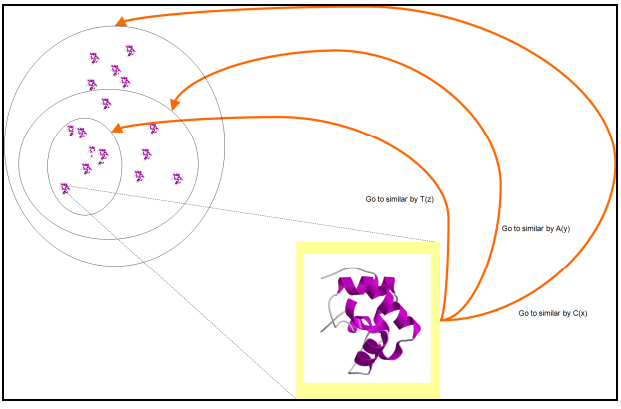

Fig. 3. Remodelling Associative Navigation

according to faceted navigation, or "trail" or "access paths", according to hypermedia design) always available to the user. A plastic representation of this conceptual remodeling is shown in Figure 2.

As each protein domain can be labeled according to its structural characteristics (each characteristic corresponds to one of the 8 CATH levels), these same characteristics can be used to design the access structures to the protein domains. The selection of a specific value of a characteristic determines a collection of protein domains, which could be browsed and accessed *as is*, independently by the other characteristics. For example, users interested in protein domain of architecture *"horse shoe"* can visualize and access *all* protein domain of with horse shoe architectures, independently by their specific topology, homologous superfamily or sequence.

As additional design opportunity, this hypertextual modeling enables supporting classic associative navigation from the protein domain. Independently by the specific access path users have chosen to reach a given protein domain, they can navigate from the details of that protein domain to other similar protein domain by architecture, topology, superfamily, or sequence. In other words, the same characteristics of the protein domains used for access purposes can be efficiently reused to offer an enriched navigational experience once a protein domain has been reached. This design can help continue the navigation through a richer exploration of the content available, thus paving the way for more serendipity in the user experience (e.g. discovering new content of interest).

5 Design and Exploratory Prototyping

On the basis of the new modeling of the navigation paradigms to be supported (on top of the existing hierarchical information architecture and content), new requirements and design alternatives has been envisioned and discussed with CATH stakeholders, usability and domain experts. An interactive prototype has been produced mainly as the basis for the discussion, as a plastic vehicle to communicate design ideas and elicit new requirements for future development.

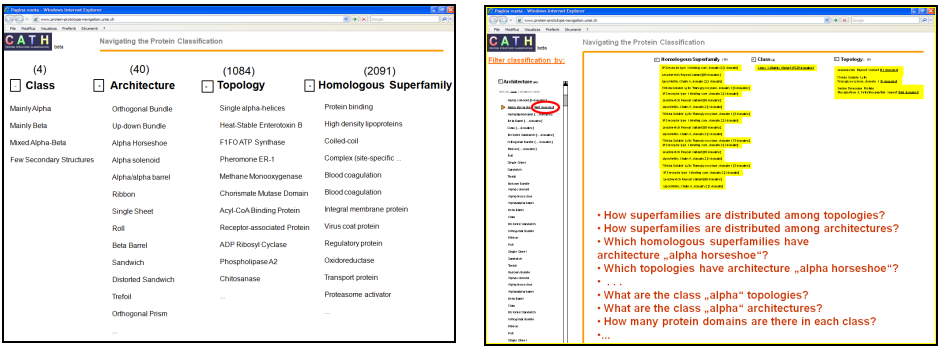

Fig. 4. Excerpts from the conceptual prototype design for advanced CATH navigation

Based on the conceptual modelling illustrated in the previous section, the main navigation console offers a semi-flat navigation structure, where users can select any CATH classification criteria to start browsing the protein domain collection. Each criterion can be expanded to include its values. For example, the criterion "architecture" includes all 40 architectures of proteins. The criterion topology includes all 1084 topologies identified, and so on. The notion of hierarchy is completely disappeared from the navigation perspective (the user can start browsing from any criterion); it remains, however, as the underlying architecture of the content.

The design concept also enables to use a classification criterion at choice as primary navigation dimension to visualize the rest of the hierarchy by the values of that criterion. For example, users can choose to select the "architecture" of the protein as primary classification criterion, select a specific architecture, for example "alpha horseshoe", and project this value over the rest of the classification, throughout all the levels down of the hierarchy. In this way, it is possible to (a) visualize the relative distribution of the cardinality of the protein domains among classification criteria; (b) efficiently skipping levels of the hierarchy; (c) get a clear idea of the cardinality of each classification being visualized. To illustrate the type of insights that can be gained by such a design, examples of possible questions that can be answered by browsing the classification are: How superfamilies are distributed among topologies? How superfamilies are distributed among architectures? Which homologous superfamilies have architecture "alpha horseshoe"? Which topologies have architecture "alpha horseshoe"? What are the class "alpha" topologies? What are the class "alpha" architectures? How many protein domains are there in each class?

Overall, the potential of the introduction of this design concept is to enable researchers to reason about the classification of protein domains, and exploring the relationships in terms of relative distribution of the protein collection, before going into the details of a specific protein data. In terms of hypermedia design, one of the cores of the user experience that this solution can support is the use of the access structures as relevant content (where insights can be gained from).

Besides supporting a more advanced navigation of the full protein collection and associative linking (Figure 5b), the new design also provides a way to superimpose multiple classifications over the previously selected protein collection (Figure 5a). This helps understand which the protein domains of a given type are within a

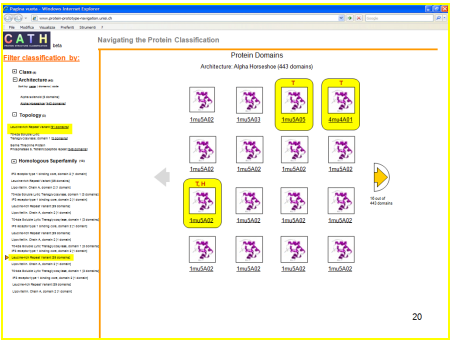

Fig. 5a. Advanced browsing in the protein collection.

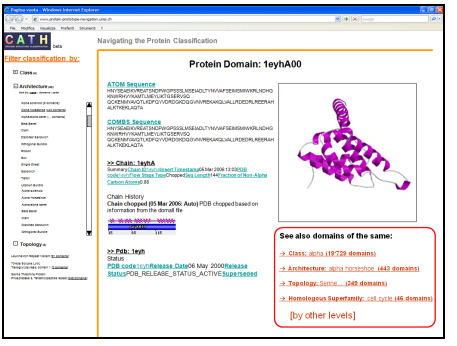

Fig. 5b. Navigating to similar proteins

previously selected collection. These and other design features enabled by the remodeling and shown by the interactive prototype are the results of three evaluation sessions: two with experts in usability, information architecture and bioinformatics and one with biologists and bioinformaticians.

6 Conclusions and Future Work

This work is an important step of a wider effort aiming at surveying a larger number of biological databases to collect and characterize the typical usability breakdowns and propose design solutions that can improve the quality of the user experience for the life science community. The ultimate goal is to make available proven design patterns (i.e. proven solutions that work) and conceptual tools in order to promote a more aware human-centered development process of bioinformatics applications. The research directions that this and other works of the authors have recently initiated in the same line open new research opportunities for the HCI and Web community to provide a rich contribution to the improvement of the user experience of bioinformatics applications, both in the area of requirements elicitation and analysis (understand the users and stakeholders involved and their goals), conceptual design (findings design solutions to effectively shape these information-intensive applications) and usability evaluation (applying and improving existing tools and techniques to cope with a domain of growing complexity).

Acknowledgements

We thank Prof. Christine Orengo and the CATH team, most notably Ian Sillitoe, for their feedback and support to the wider project work aimed at exploring usability issues in CATH. This work has been funded by a grant from the Swiss National Science Foundation (SNSF) and the UCL Experimental Cancer Medicine Centre supported by Cancer Research UK and Department of Health. Anthony Finkelstein has been supported by Cancer Research UK and the National Cancer Research Informatics Initiative.

References

1. Bartlett, J.C., Toms, E.G.: Developing a protocol for bioinformatics analysis: an integrated information behavior and task analysis approach. Journal of the American Society for Information Science and Technology 56(5), 469–482 (2005)
2. Bolchini, D., Paolini, P.: Interactive Dialogue Model: a Design Technique for Multi-Channel Applications. IEEE Transactions on Multimedia 8(3), 529–541 (2006)
3. Bolchini, D., Garzotto, F.: Quality of Web Usability Evaluation Methods: an Empirical Study on MILE+. In: Proceedings of WISE (Web Information Systems Engineering), Lille, France. Workshop on Web Usability and Accessibility. Springer, Heidelberg (2007)
4. Hochheiser, H., Baehrecke, E.H., Mount, S.M., Shneiderman, B.: Dynamic Querying for Pattern Identification in Microarray and Genomic Data. In: Proceedings of IEEE International Conference on Multimedia and Expo (2003)
5. Javahery, H., Seffah, A., Krishnan, S.: Beyond Power: Making Bioinformatics Tools User-Centric. Communications of the ACM - Special Issue on Bio-Informatics 47(11), 58–63 (2004)
6. Bolchini, D., Finkelstein, A., Perrone, V., Nagl, S.: Better Bioinformatics through Usability Analysis. Bioinformatics 25(3), 406–412 (2009) doi:10.1093/bioinformatics/btn633
7. Broughton, V.: Faceted classification as a basis for knowledge organization in a digital environment. The New Review of Hypermedia and Multimedia, 67–102 (2001)