DIA2: Web-based Cyberinfrastructure for Visual Analysis of Funding Portfolios

Krishna Madhavan, *Member, IEEE*, Niklas Elmqvist, *Senior Member, IEEE*, Mihaela Vorvoreanu, Xin Chen, *Student Member, IEEE*, Yuetling Wong, Hanjun Xian, Zhihua Dong, Aditya Johri

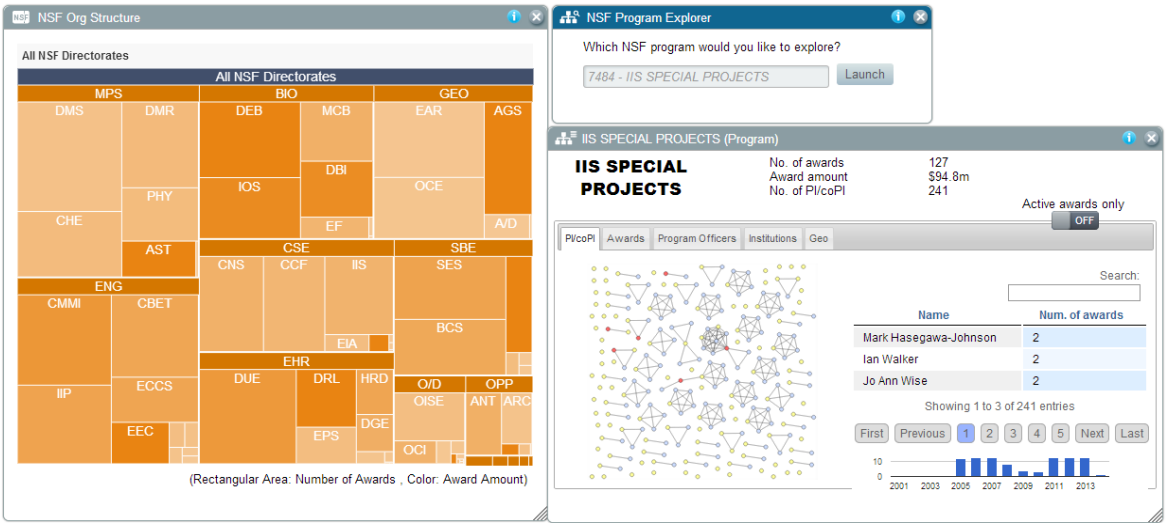

Fig. 1. Web-based dashboard showing the organizational structure of the U.S. National Science Foundation (NSF) using a treemap (left) as well as a detail view on the NSF program "IIS Special Projects" (right) in the DIA2 platform.

Abstract—We present a design study of the Deep Insights Anywhere, Anytime (DIA2) platform, a web-based visual analytics system that allows program managers and academic staff at the U.S. National Science Foundation to search, view, and analyze their research funding portfolio. The goal of this system is to facilitate users' understanding of both past and currently active research awards in order to make more informed decisions of their future funding. This user group is characterized by high expertise yet not necessarily high literacy in visualization and visual analytics—they are essentially "casual experts"—and thus require careful visual and information design, including adhering to user experience standards, providing a self-instructive interface, and progressively refining visualizations to minimize complexity. We discuss the challenges of designing a system for "casual experts" and highlight how we addressed this issue by modeling the organizational structure and workflows of the NSF within our system. We discuss each stage of the design process, starting with formative interviews, participatory design, prototypes, and finally live deployments and evaluation with stakeholders.

Index Terms—visual analytics, portfolio mining, web-based visualization, casual visualization, design study.

1 Introduction

As visual analytics technologies gain widespread adoption across a broad array of disciplines, the visual analytics community increasingly finds itself catering to an entirely new brand of users. One such population is highly-qualified professionals that are experts in their fields, yet possess little knowledge of visualization and visual analytics. Their dynamic work environment also leaves them with little time or opportunity to learn new systems. Unlike the previously proposed definition of casual visualization, which provides visualization to casual users driven by personal goals and motivations [27], we call this new brand of users "casual experts" given their extensive expertise in a domain but a casual approach to visual analytics methods. We believe that the visualization needs of such users can be best met by a design study approach [30] that investigates and understands their approach to problem solving in their domain of expertise and adapting that to the design of the visualization. We want to emphasize though that the word 'casual' in "casual expert" refers only to users' attitude towards visualizations and not their domain work which is often very high-stakes.

In this paper, we present a design study of a *web-based visual analytics* platform called **DIA2** (Deep Insights Anytime, Anywhere) designed for this new brand of casual experts. The DIA2 system is a knowledge mining platform for portfolio management [17] for STEM (science, technology, engineering, and mathematics) awards from the U.S. National Science Foundation (NSF), allowing program managers and professional staff at the NSF to view and analyze the projects, publications, and people involved in past and currently active NSF awards. The intended audience of DIA2 perfectly embodies the casual experts moniker discussed above: DIA2 users are academics with a high degree of training in their discipline, yet with little to no training and interest in advanced visualization and analytics. In keeping with the spirit of such casual experts, the design philosophy of the DIA2 project is "no manuals, no training." Instead, any training necessary in using DIA2 is designed to happen during the user experience in performing the intended tasks through several mechanisms: (1) strict adherence to norms and standards used in graphical user interface design, (2) clear labelling and a self-

- *Krishna Madhavan, Niklas Elmqvist, Mihaela Vorvoreanu, Xin Chen, Yuetling Wong, and Zhihua Dong are with Purdue University. E-mail: {cm, elm, mihaela, chen654, wong64, dong17}@purdue.edu.*
- *Hanjun Xian is with Microsoft. E-mail: hxian@microsoft.com.*
- *Aditya Johri is with George Mason University. E-mail: ajohri3@gmu.edu.*

- *Submitted to IEEE VAST 2014. Please do not redistribute.*

For information on obtaining reprints of this article, please send *e-mail to: tvcg@computer.org.*

instructive visual language, and (3) *progressive refinement* of all visual representations where visual complexity is only gradually added in response to direct and reversible actions performed by the user. The intention with the progressive refinement mechanism is that every new visual state for a particular visualization, including the first one, should be easily comprehensible to a casual expert. To convey complex data, the user would iteratively and interactively query the visualization to gradually add this complexity.

The DIA2 system (the public version of the website is accessible at http://www.dia2.org/) is a web-based interface for a large-scale online database and uses a visual dashboard design (Fig. 1). Users create data widgets on the dashboard canvas, and widgets can then be freely moved, resized, and deleted. Dashboards are persistent across sessions, and users can create and name several dashboards for different purposes. Each data widget is interactive and combines visual representations and underlying data tables for different purposes. DIA2 currently supports widgets for exploring the NSF organizational structure, concepts and keywords, investigators, institutions, research programs, and research topics. The visual representations used include treemaps, mosaic plots, ego-networks, and various statistical graphics such as bar charts, pie charts, and time-series plots, all of them implemented using the progressive refinement design guideline discussed above. Furthermore, the system supports advanced search features tying the widgets together.

The primary contribution of this design study is the participatory design process, akin to that of Sedlmair et al. [30], which we followed in creating the DIA2 platform. The design process started with closely modelling the organizational structure, workflows, and functions of the National Science Foundation. This was mainly achieved through a comprehensive set of focus group sessions and individual interviews with program officers (POs) and science assistants (SAs) at the NSF. These interviews led us to derive the concept of casual experts. We then deployed a live version of the DIA2 system internally at the NSF and evaluated the platform with members of our user groups. In this paper, we report on every stage of the research project and review results from our evaluation studies. We close the paper with a discussion and our plans for future work.

2 Background

Our work in this paper combines ideas from scientometric visualization, visual analytics for organizations, and new ideas on design study methodology. We review these research topics below.

2.1 Visualizing Research

Scientometrics is the study of measuring, analyzing and discovering science growth, structure, interrelationships and productivity [15]. It has overlapping interests with bibliometrics and informetrics. As a result, scientometrics research is often done using bibliographic visualization tools. These tools include BIVTECI [19], a prototype system proposing a minimum set of functions necessary for effective bibliography visualization, Butterfly [16], a system providing a 3D information visualizer for assessing DIALOG's Science Citation database using a so-called "organic user interface," and CiteSpace II [5], which visualizes co-authorship and co-citation relationships.

DIA2 is an analytics platform for searching, viewing, and analyzing the NSF research portfolio for casual experts. It has many features in common with the scientometrics and bibliometrics research such as most of the data is related to scientific awards, research and publications, the personal collaboration network is similar to the co-authorships in bibliometrics, and they both represent and predict cutting-edge research trends. Therefore, such bibliographic visualization techniques can also be utilized in DIA2.

A particularly relevant effort is the MultiNode-Explorer [9], a visual analytics framework that generates web-based multimodal graph visualization from multidimensional data. It accepts an entity-relationship schema transformed from the multidimensional data, a set of relational data tables, and an interface specification file, and visualizes the data as node-link diagrams. As the NSF portfolio datasets are multidimensional and multivariate, the MultiNode-Explorer framework is a useful reference implementation for our visualization process, with the important caveat that DIA2 needs additional visual representations for its multifaceted and multidimensional datasets rather than just node-link diagrams.

2.2 Visual Analytics for Organizations

Several papers in HCI have documented the obstacles encountered by large companies when conducting interface design, evaluation, and usability testing (e.g., inability of interface designers to obtain access to users, resistance to iterative design, and lack of communication) [2, 14, 26]. This previous work mainly focuses on designing products for customers rather than building visual analytics tools for internal employees. Sedlmair et al. [29] extend this work by documenting the challenges encountered by visualization researchers when designing for internal employees of large companies. They point out that the workflow, bureaucracy, and hierarchical structures could all pose challenges to the design and evaluation process. All the above-mentioned studies happen in industry settings. In this paper, we are designing for a federal government research funding agency: the U.S. National Science Foundation. NSF has very different work practices and culture compared with industrial companies, and the problems that workers need to address in their everyday work are unique to this context. Yet, common across settings is the restricted mental capacity of users to be able to pay attention to information, including visualizations.

As Green, Ribarsky & Fisher [11, 12] argue, there is general agreement among visual analytics scholars that humans are parsimonious problem solvers. As a consequence, they frequently choose the simplest heuristics that are available to them and are adequate for a given task. Therefore, by presenting information to users within a relevant context, visual analytics designers can mitigate the problem of cognitive load. In particular, for visualizations that are complex and contain numerous semantic data points, being able to leverage existing heuristics or mental models is a distinct advantage. Tory and Möller [32] in a review paper suggest that human factors are often neglected in visualization systems and argue that "more attention should be paid to users who must view and manipulate the data because how humans perceive, think about, and interact with images will affect their understanding of information presented visually. As a result, there is a strong need to study human factors as a basis for visualization design." Furthermore, they review numerous systems and show that the primary focus of system designed is on how to visually represent data to enhance data analysis but there is a lack of focus on how to visually display users' mental models and helping users improve their mental models by finding supporting and contradictory evidence for their hypotheses. They suggest that another area where visual systems can help users is providing ways to organize and share ideas. Overall, they emphasize that the visualization community needs to pay more attention to human factors-based design, specifically, how to utilize perceptual and cognitive theories. Liu & Stasko [23] look specifically at the role of mental models on visualizations and argue that although there has been some emphasis within the field on internal cognitive mechanisms, there is a need to account for ecological and situated accounts of cognitive behavior. Human cognitive functioning cannot be explained solely through description of internal mechanisms and there is always an interaction between external representations and internal representations. They review the broad literature on mental models and provide the below definition of mental models in the context of InfoVis (pg. 1001):

A mental model is a functional analogue representation to an external interactive visualization system with the following characteristics:

- The structural and behavioral properties of external systems are preserved in mental models.
- A mental model can preserve schematic, semantic or item specific information about the underlying data.

- Given a problem, a mental model of an interactive visualization can be constructed and simulated in working memory for reasoning.

2.3 Design Study Methodology

Sedlmair et al. define a design study as "a project in which visualization researchers analyze a specific real-world problem faced by domain experts, design a visualization system that supports solving this problem, validate the design, and reflect about lessons learned in order to refine visualization design guidelines" [30]. Design studies do not seek to create new visualizations; rather, they seek to solve real-word problems and provide transferable guidelines on solving such problems through reflection. Compared to technique-driven visualization research, design studies are one approach of problem-driven research. Although many design study papers have appeared in recent years (e.g. [10, 25, 28, 34]), studies that design visual analytic systems for organizations such as a government funding agency are still rare and are therefore highly valuable to the design study knowledge pool.

3 Context: U.S. National Science Foundation

The U.S. National Science Foundation (NSF) is an independent federal agency with a total workforce of about 2,100 at its Arlington, VA, headquarters. This includes approximately 1,400 career employees, 200 scientists from research institutions on temporary duty, 450 contract workers, and the staff of the National Science Board (NSB) office and the Office of the Inspector General. The NSF leadership has two major components: a director who oversees NSF staff and management responsible for program creation and administration, merit review, planning, budget and day-to-day operations; and a 24-member NSB of eminent individuals that meets six times a year to establish the overall policies of the foundation. The director and all Board members serve six year terms. Each of them, as well as the NSF deputy director, is appointed by the President of the United States and confirmed by the U.S. Senate.

NSF was created by Congress in 1950 "to promote the progress of science; to advance the national health, prosperity, and welfare; to secure the national defense…." With an annual budget of about $7.0 billion (FY 2012), NSF supports approximately 20 percent of all federally supported basic research conducted by America's colleges and universities. In many fields such as mathematics, computer science and the social sciences, NSF is the major source of federal funding for researchers and educators. NSF works to ensure that research is fully integrated with education to support the training of tomorrow's scientific and engineering workforce. According to NSF itself, it operates in a "bottom up" manner by keeping track of current research and by maintaining constant contact with the research community to keep abreast of the latest ideas and choosing the most promising people to conduct the research.

Each year NSF receives approximately 40,000 proposals of which approximately 11,000 are funded. Program officers working at NSF are responsible for the selection of proposals with the highest merit and they utilize 'review panels' to conduct proposal reviews. In order to be able to put together the panel with the right expertise, they need information about other experts in the field; they need to figure out conflicts among the proposal author and the panelists, if any; and, they need to understand the importance of an idea for the field beyond the review provided by experts, in particular to avoid duplicate funding. All these tasks require significant knowledge as well as the ability to quickly gather new information from existing data. This is also the primary need we address through our system.

4 Methods

As we approached this project, our focus was on gaining a solid understanding of users' goals, needs and workflows, which would help us ascertain their mental models. We went into the design project with a "blank slate" attitude ready to learn as much as we could about our users before creating any solutions.

To accomplish this goal, we followed Cooper's [6] goal-directed design methodology. We gained access inside the NSF and conducted nine focus groups over two separate visits with 31 NSF personnel that resulted in about eight hours of audio-recording. We analyzed the qualitative data using the method recommended by Cooper that seeks to identify similar behavior patterns that form the basis for creating personas. Three personas, described in the next section, emerged from the initial user research. We then selected a primary persona to design for. The primary persona was chosen so that any design solution that would satisfy this persona would also serve the other two.

In our case, the primary persona was a rotator. Rotators are temporary program officers who serve at the NSF for a period of time limited to two years. We then developed use cases for this persona with the help of a team member who had served as a rotator at NSF in the past. The use cases were guided by the question of what a new NSF employee would want to know in order to get up to speed with his or her portfolio of awards. We created lists of types of information the rotator would need to see and then brainstormed solutions for representing this information visually in ways that are easy to learn and understand for our casual expert users. Members of the user experience team then translated the sketches from the brainstorming sessions into detailed wireframes. The wireframes specified the layout, display, and functioning of each visualization. Special attention was paid to usability guidelines such as Nielsen's [20] 10 heuristics and Norman's concept of affordances [21]. Clearly communicating affordance, or the action enabled by each element in the design, was considered key to creating an interface that would be easy to learn. The various visualization tools were integrated under a dashboard metaphor reminiscent of financial investment dashboards – an idea that emerged from the users' frequent mentioning of the need to get a bird's eye view of their funding portfolios. During the technical implementation phase, members of both the user experience and technical DIA2 teams conducted several design reviews and cognitive walkthroughs [13] to identify and fix usability issues before launching the tool. An Alpha version was then made available to users and was evaluated using one-on-one moderated usability interviews. The results from both the formative and evaluative user research are presented next.

4.1 User Personas

Research with users inside the NSF revealed categories of users whose existence we were not even aware of. As we went into the research, we assumed program officers would be the main user group. However, three different user groups emerged from the ethnographic interviews we conducted inside the NSF. We created one persona for each user group. Cooper [6] defines personas as "composite archetypes based on behavioral data gathered from actual users" (p. 76). Personas are useful design tools because they can help designers "develop an understanding of our users' goals in specific contexts" (p. 76) as opposed to an abstract understanding facilitated by impersonal demographic information. Personas usually have a name, a photo, an explanation of the person's goals, work context, as well as needs and frustrations related to the aspect of work we design for. Three personas emerged from our formative user research:

- **James - Program Officer (PO).** James' main responsibilities are to oversee and manage research funding. He is involved in authoring calls for proposals, organizes review panels that evaluate submitted proposals, and oversees funded projects. He is often asked to prepare reports about the state of funding and relies on science assistants to find and analyze the needed data. James has a PhD in his discipline and joined the NSF because he is committed to advancing research in his discipline. However, he finds that most of his work day is spent in "fire-fighting" tasks that leave insufficient time for reflection on the broad research directions of his discipline.

- **Amy - Science Assistant (SA).** Amy has recently graduated with her MS degree. She is employed by NSF for a period limited to two years to help POs directly with data retrieval and analysis related to numerous aspects of their work. Even though Amy is highly qualified, she spends most of her day acting like a human search engine, parsing search results from databases that are difficult to query manually. It might take Amy as long as two weeks to create a report to her assigned program officer, and she cannot do so without a lot of manual work and help from other science assistants and program officers who need to review and validate her query results before she can compile any data from them. The team did not know about the existence of science assistants before conducting formative user research.
- **Matt – Rotator**. Matt is a recently tenured associate professor who is serving as a temporary program officer at NSF for a period of two years. NSF employs rotators on a regular basis as part of the organization's "bottom-up" philosophy. Matt's biggest challenge is to gain an understanding of the funding portfolio he has inherited and is now in charge of managing. It takes him months to understand the nature of the awards in his portfolio before he can become fully informed and productive.

We identified Matt as the primary persona who needed the most help. Permanent POs benefited from historical knowledge and a rich social network of colleagues they could ask for information. Rotators, on the other hand, were new to the organization, highly capable, motivated, and eager to make a difference, yet experienced huge barriers to becoming productive members of the organization. Based on the understanding of our users, the formative research ended with a list of design requirements.

4.2 Formative Design: Casual Experts

The term that emerged to describe all of our user groups was "casual experts." All three personas had advanced domain-specific expertise, but, with the exception of a few science assistants, little expertise in information retrieval and no expertise in information visualization. Moreover, they had little time and inclination to learn new visual analytics and visualization tools. It became clear that if the new system we designed required training, it would not be used. Users' mental models were all heavily influenced by the NSF's organizational and financial structure. Science assistants and permanent program officers, but not rotators, exhibited expert understanding of the organization's structure and how that influences the organization's operations and reports. The design requirements that emerged from the formative design had therefore to take into consideration the users' needs to access and assess information at a glance, while keeping it within the strict boundaries of NSF's organizational structure which was heavily reflected in users' mental models.

Getting an overview of funding portfolios emerged as the main design requirement. Users in all three groups emphasized the need to see, at a glance, how their organizational unit's funds were invested.

A second design requirement was to follow rigidly the NSF's organizational structure. It became apparent that users' mental models reflected the organizational structure. We identified the types of information needs and reports that users needed to generate (e.g. funding rates) and noticed that each one of them was dependent upon a specific organizational unit such as program, division, or code.

Third, it became apparent that users' mental models also included internal organizational language that was used rigidly and very specifically. We made an effort to learn this language and apply it to labels on the interface we designed. For example, users inside the NSF differentiate between proposals and awards, and define "awardee" as an institution, not an individual. Even though the language, organization, and work culture of the NSF were initially foreign to us, we made an effort to learn them quickly, represent them in the system design and user interface, and improve them based on continuous user feedback.

When looking at the work practices of our users, several other aspects stood out that further influenced the design of DIA2. For example, we were confronted with the array of systems that POs and SAs had to use to be able to get the required information. The organization–NSF–has over the years acquired numerous information systems that do not necessarily interact or integrate with each other. Therefore, there is no one 'place' to go to in order to find solutions to a problem. Furthermore, many of these systems might have access to the data but more crucially–from the standpoint of our design–these systems were not designed with the NSF POs and SAs in mind. The systems are largely software packages available commercially or designed in-house by contractors who primarily drew on their experience in the business community. They are not tailored towards the users from a visual analytics and visualization perspective. Even though the systems had the required information or data and could provide them to the users, the presentation was not designed optimally and resulted in little to no use by personnel who were not specifically trained to understand those systems. Furthermore, there was a persistent gap between the expertise of the users–POs and SAs–and the system designers (who were software developers with no understanding of the research context in which POs and SAs operate) that resulted in systems that were hard to use. There is another complication in that many of the POs are at NSF in a temporary position–at loan from their home institutions for a period of 2-3 years–and have to learn the numerous systems in order to be able to complete their work. When they leave, new people have to be trained. Therefore, having a useful system that reflects the 'mental model' [23] of users is a necessity.

4.3 Integrating with Existing Databases

Over the years, NSF has built several databases and systems, each containing slightly different fields for the same data. The schemas of these databases are not well documented or easily readable. The science assistants use a relatively complete Microsoft SQL database that connects to a financial system. However, due to confidentiality concerns, our project members were only given access to two less complete databases—a SQL database and a series of XML data files.

Not having access to the confidential data, we spent a considerable amount of time understanding and bridging the two available databases to generate comparable results to the database in use by the science assistants. It could be an easy pitfall for outside designers and researchers who are alien to the NSF environment to assume that one award is awarded with some amount of money, by one NSF program, under a certain topic, at a certain time. However, while working closely with SAs, we got to understand the complex data fields related to awards and proposals, including supplement, amendment, continuing grants, co-funding, sub-awards, PI transfer, the many different time stamp fields, and other complexities of managing an awards portfolio. If we had not worked so closely with our intended users, understood their work context, and done benchmark comparisons with their database, we might have run into a situation where we assumed an over-simplified data schema and designed visualizations that are not suitable for the real data.

5 Deep Insights Anywhere, Anytime (DIA2)

DIA2 is a web-based cyberinfrastructure for managing research funding portfolios for the NSF. Given that DIA2 is designed for "casual experts", the system architecture of DIA2 shields the end-users from the technical complexity of managing data and visualizations. DIA2 utilizes an n-tier architecture [18] that treats every layer of the system as a service provisioned to the other layers of the system. Therefore, DIA2 also exhibits all the properties of a service-oriented architecture [3, 8]. DIA2's technical core is designed to ease maintenance and increase availability of the system.

The system stack is divided into 4 individual nodes – with 2 nodes dedicated for the data layer (also search), 1 machine hosting the middleware components, and the final node hosting the user facing web server. In addition to these 4 nodes, an extra node serves as a fail-over node. Furthermore, the DIA2 system also accesses a scrap data storage system as needed, particularly during data

acquisition and processing. The size of scrap data storage ranges from 4 to 10 terabytes depending on the transactions in progress. The entire system stack is replicated 3 times – with the first replication set serving as the production environment, the second set providing all quality assurance services, and the final set acting as the development environment. End-users interact only with the production nodes. All aspects of DIA2 are optimized to decrease the time required to provide end-users with appropriate insights. To this end, DIA2 utilizes a hybrid of traditional disk drives spinning at 5,400 rpm and solid state drives (SSDs). The traditional disks house all middleware components and non-data entities, while the SSDs contain all databases, search indices, and query caches.

5.1 Data Layer, Search Indices, and Query Caches

DIA2 uses a combination of structured and unstructured data entities as the base for all the analytic services provided to the end users. The primary database system is a MySQL database that consists of a variety of metadata relevant to the grants that have been made by the NSF. In addition, DIA2 also uses full texts of awards abstracts, journal papers and conference proceedings resulting from a sizeable number of awards, and in many cases actual links harvested from focused crawling [4, 7] of the web for products resulting from the NSF awards (such as curricular materials and web resources). DIA2 also includes a warehouse of data derived from surveys conducted as part of NSF program analyses, impact reports generated by the individual NSF programs, and also taxonomies developed by individual programs within the NSF.

The data layer includes a range of acquisition, aggregation, disambiguation, and completion protocols that ensure data coverage and data cleanliness. Given that DIA2 users are "casual experts", the types of insights they require also demand a high level of precision. Many times users of DIA2 are responding to congressional requests for information or other high stakes decision-making contexts. To this end, the data layer includes a set of protocols implemented via system daemons to continuously evaluate the quality of the data, incrementally request additional data from the various systems inside the NSF, and resolve ambiguity in author names, proposal titles, institutional affiliation, and so on. For example, an awardee John Smith maybe affiliated with Institution A when he first receives an NSF award. Over time, John Smith may have received supplemental funding for this award. As John Smith is a highly successful researcher, he may move to another Institution B and move his award along with him. DIA2's data layer has to systematically track all of these scenarios and then account for the amount of money moved and also be careful as to not double count the award information. The problem is complicated when there is another John Smith at Institution B who works in a completely different field who may have received research funding from the NSF. DIA2 includes carefully designed data tracking methods to account for these types of ambiguities. Furthermore, DIA2 executes these protocols on a continuous basis to keep data clean and complete. Currently DIA2 archives data from the year 1973 to March 2014. We have strategically decided (based on user studies) to expose only data from 1995 to increase the utility of the system to our end users.

Search is a critical feature within DIA2. Users can search for a program element, a person, an institution, or any keyphrase within the database. It is critical to point out that while users search for an entity, the results are always synthesized into data dashboards (using appropriate widgets) and are never returned to the users as a single list, which is common with most commercial search engines today. We elaborate more on this in a later section. Within the system, DIA2 utilizes 3 different methodologies to provide end users with a refined search experience: (1) For every document, DIA2 systematically extracts as set of keyphrases that best describe the document. Keywords denote a single word and a keyphrase denotes multi-word units. Keyphrases are valuable in describing the content of single documents and provide a kind of semantic metadata and document summary that is useful for a wide variety of purposes. As large document collections such as digital libraries become widespread, the value of such summary information increases. Keywords and keyphrase are particularly useful because they can be interpreted individually and independently of each other. They can be used in information retrieval systems as descriptions of the documents returned by a query, as the basis for search indexes, as a way of browsing a collection, and as a document clustering technique [33]. (2) DIA2 allows folksonomic tagging of documents and data entities – meaning developing a search taxonomy based on user supplied keyphrases. Folksonomies developed via user input can be extremely valuable in identifying and distinguishing between documents with a high degree of confidence. Given the high stakes nature of the queries that users may execute on DIA2, this methodology offers an extremely effective seeding for search. (3) Finally, DIA2 utilizes Apache Solr[1] (a derivate of Apache Lucene[2]) to index a wide range of documents. Solr allows DIA2 to grow its indices to an extremely large scale (to the tune of many petabytes). Each time data is acquired by DIA2 and post processing is complete, DIA2 automatically triggers a system daemon to prepare an appropriate XML file that serves as input to Solr, which in turn indexes the data as required.

When users interact with DIA2, every user click produces a huge demand for data. Traditionally, clicks trigger requests to the database creating a high possibility of a bottleneck at peak data demand. However, in DIA2 every user click automatically routes the request to the Query Cache Handler (QCH) – which immediately checks the Query Cache to evaluate if any previous user has processed the same or similar requested result. If a positive match is found, the QCH immediately returns the results and no further traffic is initiated towards the database. If a query is triggered for the very first time, the QCH automatically keeps a copy of the results returned with an appropriate timestamp. Given the complexity of the views provided to the end users, the QCH cuts down response times of complex queries by nearly 90%. Many data widgets found within the DIA2 user front-end rely heavily on the Query Caches to perform their tasks effectively. For example, the widgets using the treemap visualizations rely heavily on the QCH to increase responsiveness. The QCH is always active even upon data acquisition. DIA2 proactively anticipates the data needs from the users and automatically caches many results that are very often requested. All of the QCH functions reside within the DIA2 data layer. It must be pointed out that the QCH is housed on the solid state drives (SSDs) to decrease responsive times due to data reads (or writes).

Finally, it must be pointed out that the entire data layer is exposed to other parts of DIA2 as a set of JSON/RPC services. There are two primary reasons for this architectural decision. (1) Security of the primary data sources is highly critical in systems such as DIA2. Traditionally, developers transmit their queries to the database engines directly without an intermediary. While this is effective in smaller systems, the JSON/RPC services allow for better consistency, maintenance, and security of the data components. (2) Secondly, experts in data mining may not want to be constrained by the UI provided to "casual expert" users and may want to work directly with the raw data for a variety of purposes. This data architecture allows appropriate rationing and control of the data flow out of the DIA2 system while providing standardized data access.

5.2 Middleware Layer

Most of the algorithms, workflow artifacts, and rules that drive various aspects of DIA2 are managed within the middleware layer. DIA2's middleware layer is designed as a set of services that are invoked as needed. As opposed to a traditional approach that invokes algorithms for various functionalities, DIA2 deploys each algorithm into a generic JSON/RPC wrapper that can be invoked on demand and is highly abstracted. The middleware is divided into three parts:

[1] http://lucene.apache.org/solr/

[2] http://lucene.apache.org

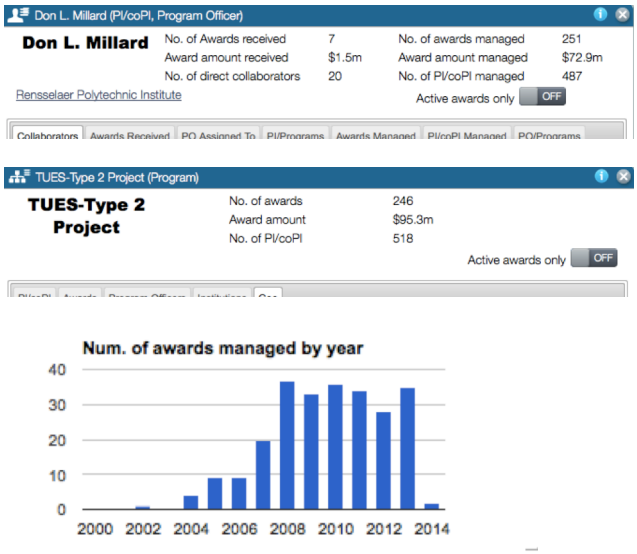

Fig. 2. The rules engine determines how the values displayed in each of the above views (presentation layer) are calculated. It aggregates information at the program level or individual –roles level.

5.2.1 Rules Engine

DIA2 provides end users with summative view of various elements in a single view. For example, DIA2 regularly identifies total award amounts, number of awardees, and total number of awards. These analytics are provided not only at various levels of organizational structure, but also for individuals. DIA2 recognizes the role of an individual and can automatically provide the statistics related to a specific role an individual plays. While these appear to be simple numbers, the computation of these values is by no means trivial. There are very specific rules on determining what counts as an award, how to determine in which year an award was made, track funding movement with change in awardee information, and determine what organizational entities were responsible for various award sub-parts. The rules engine also has very specific date calculations that calculate when an award is active. Fig. 2 shows the rules engine driving the analytics provided to the end-users. Furthermore, the rules engine also tracks and provides appropriate processing frameworks to other DIA2 visualization services (such as the treemap visualization) and search aggregation.

5.2.2 Visualization Services

While the rules engine discussed in the previous section determines the appropriate data aggregation and processing framework, visualization services work in coordination with the presentation layer (discussed in the next section) to render the appropriate visual information requested by the end users. DIA2 design considers visualization of information from a highly utilitarian perspective. The methodology used to determine the affordance that a specific visualization provides to the end users is discussed in [24]. As opposed to thinking of individual visualizations as algorithms, DIA2 considers the basic nature of the data to be visualized and creates a service that is generic and abstract enough to serve visualizations specific to data types. We elaborate on this next.

Hierarchical Data Visualization Service (HDVS): DIA2 users are very interested in data that reflect hierarchy. The preferred visual representation of hierarchical data within DIA2 is through the use of treemaps [31]. This service provides all of the processing needed for processing organizational structure, programmatic structure, and also taxonomy information within DIA2. Fig. 3 shows the variety of hierarchical data that is processed within DIA2. This single service provides the algorithmic core for all of the information processing presented in Fig. 3.

Collaboration Data Visualization Service (CDVS): One of the purposes of DIA2 is to showcase the collaboration networks emerging around individual researchers and also around organizational entities. To showcase the collaboration around individual researchers, DIA2 utilizes ego-centric social networks, while the organizational structures are visualized using simple flat spring loaded social network layout. Fig. 4 highlights the type of visualizations provided by the CDVS.

Geographical Data Consolidation Service (GDCS): DIA2 is at its very core a portfolio mining platform. Evaluation of how federal funding is distributed across geographical area is a critical part of the analytics that needs to be provided to the end users. During the initial requirements gathering phase of DIA2, end users repeatedly emphasised the need for services that allow aggregation of data across geographical regions. To this end, DIA2 is capable of not only providing consolidated data on map overlays, but also drill down into data aggregation at the level of individual academic institutions focused within a specific geographical region. Fig. 5 provides a simple example of the GDCS in action. The GDCS also allows quick comparisons of various data aggregations across the geographical range.

5.2.3 Search Services

One of the very unique capabilities of DIA2 is its ability to translate any search into a coherent set of analytics. Within DIA2 all data artifacts – people, organizational structures, programs, awards, concepts, keyphrases, and institutions – are searchable. However, the search services are designed to not return a linear list of results. The aggregated search service continuously interacts with the rules

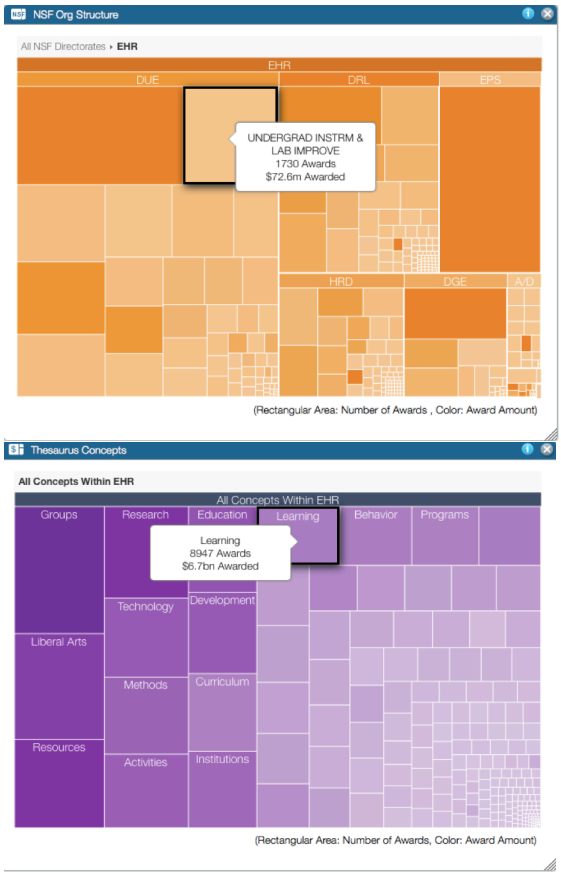

Fig. 3. DIA2 presents a range of organizational and thesaurus data through hierarchical views using treemaps. In this figure we show the NSF org view, a program view, and a thesaurus view all using a single service.

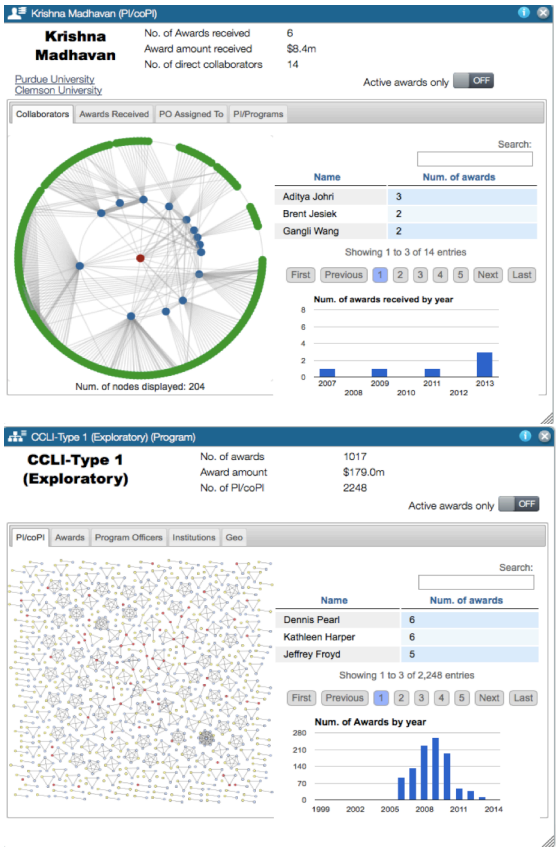

Fig. 4. DIA2 includes a range of services that allow better visual presentation of collaboration data. Nodes represent researchers and links are proposals that have been awarded by NSF. These visualizations also show capacity building within each program (organizational structure).

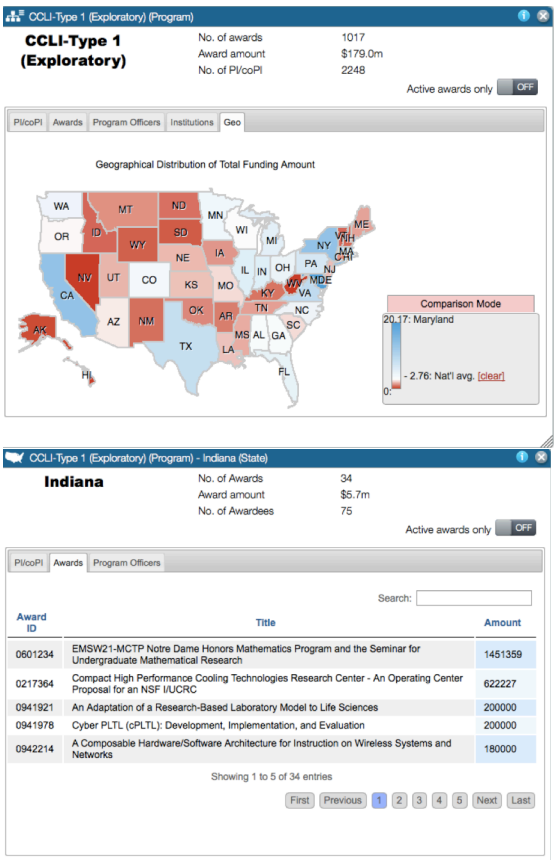

Fig. 5. DIA2 provides a full range of analytics focused on geographical locations. The first graph in the inset is the GDCS working in the "Comparison Mode".

engine and the visualization services to synthesize the results in meaningful ways. Fig. 6 shows a simple search of a concept driving a full set of highly synthesized results.

The search results are provided in a simple widget that contains multiple tabs. Each tab has information relevant to one aspect of the search. The search service allows end users to use a variety of operators such as AND, OR, and NOT to constrain the search results effectively. Furthermore, the search service is also evolving to provide users with the ability to define abstract concepts through a folksonomy methodology. For example, users could define the term "cyberlearning" using a set of vectors such as technology-enhanced learning, game-based learning, and mobile learning. DIA2 will track this user-supplied definition and automatically search for the entire search vector each time cyberlearning is searched for. This grouped search service will be introduced in the coming months.

5.3 Presentation Layer

The previous sections provided a description of the data and middleware layers respectively. However, these layers are completely hidden from end users. The only aspect of DIA2 that users really interact with is the presentation layer. The entire user experience within DIA2 is based on a dashboard metaphor. Users are provided with 3 dashboards (blank canvases) by default with the option to add up to 5 dashboards in any workspace. The limits were determined based on a simulation of resource allocation to enable scaling to a large number of users. All dashboards can be named and saved for future use. In future versions of DIA2, dashboards are also designed to be shareable with other users. Each dashboard can be composed of multiple widgets. Currently DIA2 supports 6 different widgets with 3 more currently being planned. Each widget provides data views with multiple tabs providing relevant information to the end users. Every aspect of the presentation layer is completely controlled using an XML configurator file. All widgets have a standard descriptor that is packaged along with the code for that particular widget. The descriptor sets and determines the behaviors of the widget and also provides a baseline for the middleware services the widget needs to connect with to provide its functionality.

As users launch the alpha version of DIA2, they are provided with a simple widget selector called the DIA2 Guide. A wizard (Fig. 7) also provides users with a brief description of each widget.

As new widgets are introduced, the DIA2 guide will automatically detect the existence of a new widget and provide the users with option to select the new widget. As widgets are selected into the workspace a small icon showing the status of the widget appears on the dashboard indicator on the tab. All dashboards can be saved and cleared. The presentation layer also includes a caching mechanism to speed the rendering of the widgets on user screens.

6 Evaluation

The alpha version of DIA2 was tested with users from the NSF using a moderated usability interview that focused on assessing ease of learning. Four program officers and two science assistants agreed to examine the interface and describe out loud their thoughts [22] as they tried to understand what the tool did and how to use it. Even though we asked participants to perform some tasks using DIA2, we chose not to collect quantitative metrics such as time on task [1] and instead to focus on users' cognitive processes for understanding of the interface. The moderated sessions were video recorded, and the recordings were analyzed in order to identify usability issues. We define a usability issue as any aspect of the interface that users did not readily understand or were unable to use.

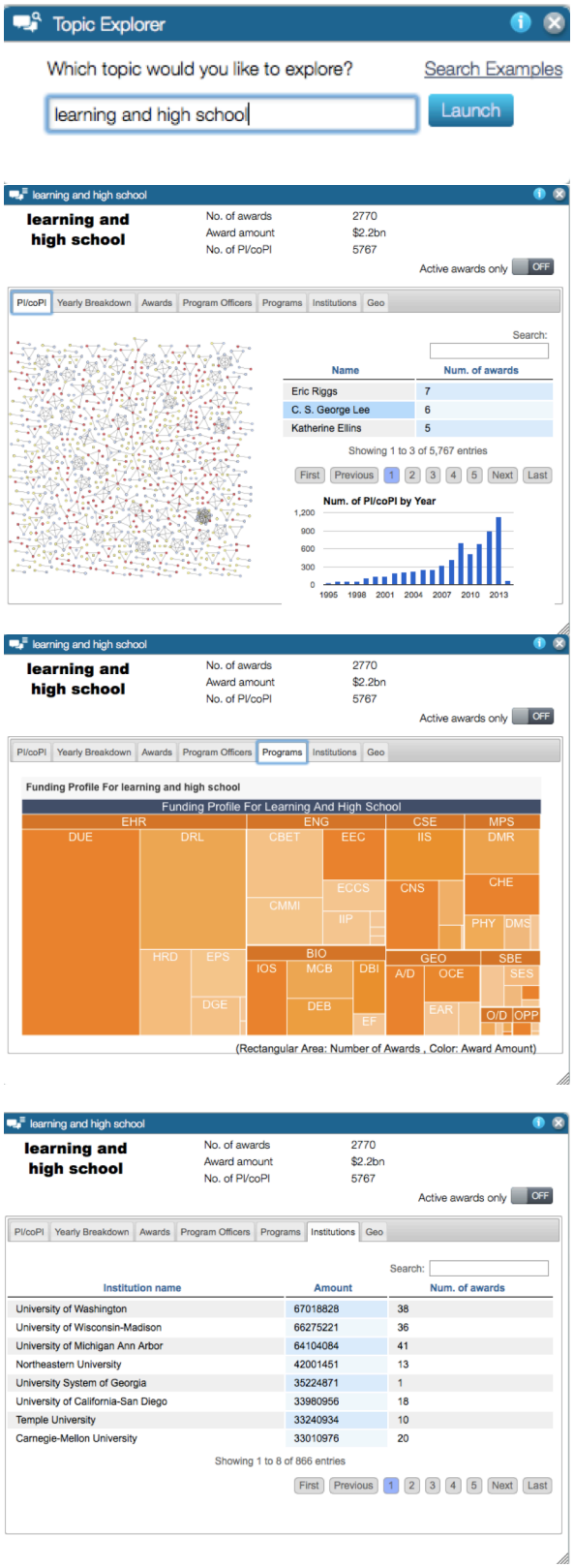

Fig. 6. Search service interacts with the rules engine and visualization services to provide end users with highly aggregated and processed results as opposed to long list of results.

Overall, the system received positive feedback from users, who made comments such as "this thing reads my mind" and "I feel it was designed for me." These are indicators that the users' mental models were indeed reflected in the design. However, feedback on specific visualizations indicated areas of improvement.

Specifically, the **treemap visualization** is used repeatedly in DIA2 to show the allocation of funds and numbers of awards (proposals) across organizational units. None of our users seemed to be familiar with treemap visualizations prior to the testing. They used their knowledge of the organization's structure to infer the meanings associated with block size and color saturation on the treemap visualization. All users were able to figure it out after a few seconds of thinking. Even though a one-line legend was available at the bottom to explain the meaning of block size and color, users did not read it. This finding is interesting from two different points of view. First, our casual experts assume sufficient expertise and enjoy figuring things out by themselves, so they are unlikely to read any instructions, no matter how short. Second, the use of organizational units they were already familiar with seemed to facilitate their quick learning of the treemap visualization.

DIA2 also uses a number of traditional data representations such as **tables and bar charts**. Even though we identified a number of usability issues related to sorting and pagination of information in tables, users, as expected, had no trouble understanding them.

Ego networks are another type of visualization frequently used in DIA2. DIA2 uses ego networks to show collaboration on NSF proposals among individuals. For example, a program officer can see not only a list of all the individuals and awards he or she has funded, but also a social network of all the individuals funded. We assumed that collaboration networks would be an interesting measure of a discipline's development and would be useful in helping program officers identify conflicts of interest. Close collaborators, for example, cannot serve on panels evaluating each other's proposals. However, such network representations were not part of our users' mental models. Labels for each node appear in the visualization only on mouse-over. The nodes are represented as small circles. Therefore, the networks failed to clearly and quickly communicate to users that the nodes were people. Upon exploration, most users understood what the networks represented, although a couple of them needed some explanations from the moderator. Even so, the users were not sure what the links between nodes represented, or what the meaning of the color-coding was. They evaluated the visualization as "cool" and "interesting" but were not yet sure as to how they would use it in their daily work.

Collectively, these findings show that the treemap visualization, which was completely new to users, made more sense to them than the relatively popular social network visualization. We explain this finding by reverting to this project's guiding concept, that of mental models. The way we used treemap visualizations in DIA2 was consistent with the users' mental model and therefore they could rely

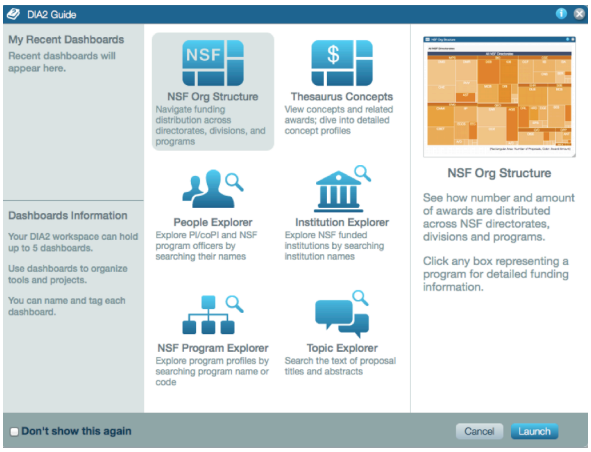

Fig. 7. DIA2 Guide available to users at the launch of the alpha version. The guide provides a simple selector that users can select to learn about the widget and also launch it easily.

on that existing knowledge to make sense of the new information. However, because NSF staff members rarely think about collaborative research networks, seeing investigators represented this way made less sense to them than the treemap. Users showed interest in this new perspective and were open to its potential, but had a hard time learning it on their own.

7 Discussion

Sedlmair et al. [30] point to the specific characteristic of design studies as using visualization expertise to understand and build solutions that are able to address real-world problems faced by domain experts. DIA2 is precisely such as system. The end-users of our system require deep insights about their portfolio at a very high level of precision to be able to address various real-world policy concerns and direct funding. To this date, as far as we know, no other team in the world has managed to derive such deep insights into the real-world problems faced by users within a governmental agency such as the U.S. National Science Foundation. Our solution not only utilizes publicly available data, but DIA2 is being deployed directly inside the NSF firewalls. This level of impact comes with significant design issues that we address next.

7.1 Designing Based on Metadata Schemas Only

One of the most important and critical reasons why previous efforts to build a portfolio analysis system by external researchers (meaning not staff, employed by, or under contract of) the NSF is that research teams cannot have access to internal datasets directly. This made building a data-driven solution virtually impossible. One of the very unique contributions of DIA2 is the realization that data-driven systems can be built as long as access to metadata schemas can be provided. The DIA2 team did not have direct access to the NSF data. In fact, we were never allowed to look at the data. However, using our design process, as described in this paper, we were still able to understand the users, derive the user requirements, build highly tailored solutions for the audience, and then deploy this solution. Using user testing and evaluation of our solutions we are designing DIA2 to address a major national need in understanding the NSF's funding portfolio. This requires close collaboration and trust between the external researchers and users inside the NSF.

DIA2 is in essence a great example of how to work within the legal framework of data at federal agencies while still delivering value using visual analytics. Working closely with our intended users (particularly the science assistants), we were able to "design in the dark" and reach the intended result without ever seeing the real data. A final critical component in our approach was adopting an agile development method of releasing early and often; our users on the other side of the wall (i.e. who had access to the confidential data) could then give us rapid feedback on the results to allow for changes.

7.2 Affordance is Innovation

The visualization and visual analytics community has at times a tendency to dismiss applied work as not innovative. The true value of visualizations or indeed visual analytics has to be in the affordances its use offers to end-users. DIA2 takes on the challenge of providing insights at speed in a context where the stakes are high. From our user studies, we understand that in the design of systems like DIA2, it is extremely critical to select simple and useful representations of data rather than to strive for the creation of absolutely new algorithms and visualizations.

Furthermore, what is even more critical is to offer insights at high speed while reducing the cognitive burdens on the end-users. For example, within DIA2 many of the visuals provided reduce the workload on end-users by tens of hours if not more. This enables them to perform many more analyses in meaningful ways than before. Also, they are now able to ask more critical questions than were possible before. It is to this end, every aspect of DIA2 is highly optimized to function at high speed, yet reliably. Our argument in this paper is that for user-focused systems like DIA2, ensuring that the end-users maximize on the value of the knowledge mining platform is far more important than novelty in the visual representations and analytics - *affordance is innovation*.

7.3 Fitting into Existing Organizational Ecosystems

One of the biggest challenges to introducing a system such as DIA2 into an environment like the U.S. NSF is that it needs to fit in with the organizational and cultural norms of that institution. Furthermore, even with the public data, the simple visuals and ability to mine massive amounts of data in an easy and intuitive way opens up the awards portfolio to a level of scrutiny that organizations need to prepare and plan for. It is true that such data are available publicly, but what is different is the ability to see the strengths and weaknesses of a program or organizational branch very simply. Furthermore, new systems such as DIA2 that are introduced into an organization must inevitably adapt to an existing ecology of both software—such as databases, management software, and search interfaces—as well as hardware—including server rooms, network architectures, and security systems. Adapting both the software and hardware aspects of DIA2 to the needs of end-users without losing on the ability to innovate scientifically is truly non-trivial.

8 Conclusion and Future Work

In this paper we have presented a case study of DIA2, a project designed to facilitate portfolio mining for the U.S. National Science Foundation program officers and assistants. In this system we have targeted a novel user population as well as a novel problem domain. Although a number of internal applications were available to users, none of them were designed with the 'user' in mind; they were standardized packages modified for the users. As a consequence, the available solutions often proved inadequate, and adapting them to the users was hard for the designers as they did not understand the domain of the users. In our effort we had to start from scratch and our initial plan was to use novel and popular techniques currently in vogue and that had proved useful for a lot of other domains (dashboards). We revised our plans and started with the requirements of the target task and related the techniques to what had come up in the interviews and in feedback. We wanted our system to be able to provide new insights to the users but also support them in their tasks and reduce the time needed to respond to questions. The visualizations are a 'palette' of different kinds that are useful for understanding this domain and similar domains where organizational structure and function are largely in silos with some integration across functions. Although our work does not contribute novel techniques or algorithms per se, the novelty of our work lies in our design approach and targeted domain. There are several design lessons learned from this case study such as how to design for specific organizational structures, and, how to translate mental models into design requirements and visualizations.

Our future work will focus on continuing to develop the DIA2 system in response to our end-users. We are also developing a community-facing version of the system that will help answer the same portfolio mining questions for our colleagues in the scientific community. Finally, we are highly interested in the concepts of casual experts and progressive refinement for visual analytics, and hope to continue exploring how to better accommodate these design constraints in future visual analytics and visualization systems.

Acknowledgment

This work was supported by the U.S. National Science Foundation awards TUES-1123108, TUES-1122609, and TUES-1123340. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] W. Albert and T. Tullis, Measuring the user experience: collecting, analyzing, and presenting usability metrics. Newnes, 2013.

[2] J. O. Bak, K. Nguyen, P. Risgaard, and J. Stage, "Obstacles to usability evaluation in practice: a survey of software development organizations," in *Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges*, 2008, pp. 23–32.

[3] M. Bichier and K.-J. Lin, "Service-oriented computing," *Computer*, vol. 39, no. 3, pp. 99–101, 2006.

[4] S. Chakrabarti, M. Van den Berg, and B. Dom, "Focused crawling: a new approach to topic-specific Web resource discovery," *Computer Networks*, vol. 31, no. 11, pp. 1623–1640, 1999.

[5] C. Chen, "Visualising semantic spaces and author co-citation networks in digital libraries," *Information processing & management*, vol. 35, no. 3, pp. 401–420, 1999.

[6] A. Cooper, R. Reimann, and D. Cronin, *About Face 3: The Essentials of Interaction Design*, 3rd ed. Wiley, 2007.

[7] M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles, and M. Gori, "Focused Crawling Using Context Graphs.," in *VLDB*, 2000, pp. 527–534.

[8] T. Erl, Service-Oriented Architecture (SOA): Concepts, Technology and Design. Upper Saddle River, NJ: Prentice Hall PTR. 2005.

[9] S. Ghani, N. Elmqvist, and D. S. Ebert, "MultiNode-Explorer: A Visual Analytics Framework for Generating Web-based Multimodal Graph Visualizations," *Proc. of EuroVA'12*, pp. 67–71, 2012.

[10] S. Ghani, B. C. Kwon, S. Lee, J. S. Yi, and N. Elmqvist, "Visual analytics for multimodal social network analysis: A design study with social scientists," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, no. 12, pp. 2032–2041, 2013.

[11] T. M. Green, W. Ribarsky, and B. Fisher, "Building and applying a human cognition model for visual analytics," *Information visualization*, vol. 8, no. 1, pp. 1–13, 2009.

[12] T. M. Green, W. Ribarsky, and B. Fisher, "Visual analytics for complex concepts using a human cognition model," in *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*, 2008, pp. 91–98.

[13] V. Grigoreanu and M. Mohanna, "Informal cognitive walkthroughs (ICW): paring down and pairing up for an agile world," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 3093–3096.

[14] J. Grudin, "Obstacles to participatory design in large product development organizations," *Participatory design: principles and practices*, pp. 99–119, 1993.

[15] W. W. Hood and C. S. Wilson, "The literature of bibliometrics, scientometrics, and informetrics," *Scientometrics*, vol. 52, no. 2, pp. 291–314, 2001.

[16] J. D. Mackinlay, R. Rao, and S. K. Card, "An organic user interface for searching citation links," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1995, pp. 67–73.

[17] K. Madhavan, M. Vorvoreanu, N. Elmqvist, A. Johri, N. Ramakrishnan, G. A. Wang, and A. F. McKenna, "Portfolio Mining," *IEEE Computer*, vol. 45, no. 10, pp. 95–99, 2012.

[18] P. D. Manuel and J. AlGhamdi, "A data-centric design for n-tier architecture," *Information Sciences*, vol. 150, no. 3, pp. 195–206, 2003.

[19] D. Modjeska, V. Tzerpos, P. Faloutsos, and M. Faloutsos, "BIVTECI: A bibliographic visualization tool," in *Proceedings of the 1996 conference of the Centre for Advanced Studies on Collaborative research*, 1996, p. 28.

[20] J. Nielsen, "10 Heuristics for User Interface Design." [Online]. Available: http://www.useit.com/papers/heuristic/heuristic_list.html.

[21] D. Norman, *The design of everyday things*. Basic books, 2002.

[22] C. Lewis, Using the" thinking-aloud" method in cognitive interface design. IBM TJ Watson Research Center, 1982.

[23] Z. Liu and J. T. Stasko, "Mental models, visual reasoning and interaction in information visualization: A top-down perspective," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 16, no. 6, pp. 999–1008, 2010.

[24] Q. Liu, M. Vorvoreanu, K. P. Madhavan, and A. F. McKenna, "Designing discovery experience for big data interaction: a case of web-based knowledge mining and interactive visualization platform," in *Design, User Experience, and Usability. Web, Mobile, and Product Design*, Springer, 2013, pp. 543–552.

[25] M. Ogawa and K.-L. Ma, "code_swarm: A design study in organic software visualization," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 15, no. 6, pp. 1097–1104, 2009.

[26] S. E. Poltrock and J. Grudin, "Organizational obstacles to interface design and development: two participant-observer studies," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 1, no. 1, pp. 52–80, 1994.

[27] Z. Pousman, J. T. Stasko, and M. Mateas, "Casual information visualization: Depictions of data in everyday life," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 13, no. 6, pp. 1145–1152, 2007.

[28] M. Sedlmair, A. Frank, T. Munzner, and A. Butz, "RelEx: Visualization for actively changing overlay network specifications," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, no. 12, pp. 2729–2738, 2012.

[29] M. Sedlmair, P. Isenberg, D. Baur, and A. Butz, "Information visualization evaluation in large companies: Challenges, experiences and recommendations," *Information Visualization*, vol. 10, no. 3, pp. 248–266, 2011.

[30] M. Sedlmair, M. Meyer, and T. Munzner, "Design study methodology: Reflections from the trenches and the stacks," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, no. 12, pp. 2431–2440, 2012.

[31] B. Shneiderman and M. Wattenberg, "Ordered treemap layouts," in *Information Visualization, IEEE Symposium on*, 2001, pp. 73–73.

[32] M. Tory and T. Moller, "Human factors in visualization research," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 10, no. 1, pp. 72–84, 2004.

[33] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, "KEA: Practical automatic keyphrase extraction," in *Proceedings of the fourth ACM conference on Digital libraries*, 1999, pp. 254–255.

[34] J. Wood, D. Badawood, J. Dykes, and A. Slingsby, "BallotMaps: Detecting name bias in alphabetically ordered ballot papers," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 12, pp. 2384–2391, 2011.