Making novel connections between literature and data

Donald A. Pellegrino Jr.

Drexel University

Philadelphia, PA USA

November 24, 2009

Contents

1 Summary **1**

2 Literature Review **1**

2.1 The Scientific Research Information Environment 2

2.1.1 Cyberscholarship, Cyberinfrastructure, Discovery and Innovation 2

2.1.2 General Digital Library Systems 3

2.1.3 Domain Specific Digital Library Systems 4

2.1.4 Open Notebook Science 4

2.1.5 Models of Scientific Communication 5

2.1.6 Information Overload 6

2.2 Navigating the Environment 7

2.2.1 Bibliometrics 7

2.2.2 PageRank 9

2.2.3 Literature Related Discovery 10

2.2.4 Data Related Discovery 11

2.2.5 Visual Analytics 13

2.2.6 Scenario Visualization 15

3 Problem Statement **16**

3.1 Problem being addressed 17

3.2 Motivation 17

3.3 Research Questions 17

3.3.1 Establishing connections 18

3.3.2 Algorithms for discovering novel connections 21

3.3.3 Extending Digital Libraries 27

4 Evaluation Strategies **28**

List of Tables

1 Doctoral Candidacy Committee iv

2 Selected NSF Solicitations 3

3 High level classifications in the UNISIST model. 8

4 Research Questions 17

List of Figures

1 The original UNISIST model as reproduced in [SAH03]. 5

2 UNISIST model updated to reflect effects of the Internet [SAH03]. 7

3 All VAST 2008 Challenge data unified into matrices [PCM$^+$08, Figure 1]. 20

4 Adding hypotheses directly to the integrated model [PCM$^+$08, Figure 7]. 21

5 Associative network of heterogeneous data [PCM$^+$08, Figure 8]. 24

6 An interesting subgraph [PCM$^+$08, Figure 10]. 25

7 Violation of domain constraints [PCM$^+$08, Figure 11]. 26

Requirements for Advancement to Candidacy

This document is submitted to the faculty of the College of Information Science and Technology, Drexel University in partial fulfillment of the requirements for the degree of Doctor of Philosophy. The requirements for advancement to candidacy in the College of Information Science and Technology are documented in the "Advancement to Candidacy" section of "The PhD Degree Program Description & Procedures." The contents of this section are quoted below [WFE08, pp.13–14]:

> To advance to candidacy a PhD student presents a critical literature review along with a problem statement in his or her area of interest. A typical problem statement should contain a description of the problem being addressed, a discussion of the importance of the problem, why it is interesting, and a set of research questions. The problem statement will usually be at least two or three pages long. The overall length of the literature review and problem statement will vary depending on the topic; a rough guideline is 15–30 pages (double-spaced).
>
> Before advancing to candidacy, the student must finish all coursework and have a successful final portfolio review. Note that the procedure for advancing to candidacy is separate and distinct from the final portfolio review. The student is expected to complete the literature review and problem statement within six months of finishing coursework.
>
> The literature review and problem statement is evaluated by the student's Doctoral Candidacy Committee, as defined in the University's doctoral policies (five members in total, including the chair and four members, with one or two members from outside IST or outside the University). The student meets with the full Committee synchronously to discuss the literature review and problem statement.
>
> The student advances to candidacy (passes the "Candidacy Exam") if the Committee approves the literature review and problem statement. If the student's literature review and problem statement is not accepted the first time, the student may revise and resubmit it to the full Committee a second time. The time limit for resubmitting is six months.
>
> Students who fail to advance to candidacy after two attempts will be removed from the doctoral program.
>
> The University allows five years for completing the doctoral degree (with the possibility of an additional extension of two years, given good progress). This limit applies to both full-time and part-time students. It is advisable for all students to advance to candidacy as soon as possible after the coursework is finished. Since part-time students usually take longer to complete their coursework, they, in particular, should consider beginning to develop the literature review and problem statement while still completing coursework.
>
> University forms that must be filed for advancement to candidacy are the following:
>
> - Form D-3 "Doctoral Candidacy Committee Appointment and Exam Schedule" must be filed with the Office of Graduate Studies four weeks before the Candidacy Exam meeting takes place.

- Forms D-4 and D-4a "Reports on Candidacy Examination" are submitted to the Office of Graduate Studies when the student's literature review and problem statement is approved by the Committee.

The composition of the five member Doctoral Candidacy Committee is listed in Table 1.

1. Chair	Chaomei Chen
2. IST Member	Xia Lin
3. IST Member	Robert Allen
4. Non-IST Member	Jean-Claude Bradley
5. Non-IST Member	Longjian Liu

Table 1: Doctoral Candidacy Committee

1 Summary

One specific mechanism of scientific discovery is the creative process of making novel connections between previously disconnected bodies of knowledge [FMC07, Swa86b]. Many tools and methods exist for providing access to the literature that defines a field. These rely on indexes of the literature and measures of its connectedness. Researchers faced with the problem of connecting relevant data and other artifacts that are not part of the literature are lacking tools to support their search. An increase in the digital availability of research artifacts has created an opportunity for them to play a larger role in the research process. For a researcher to make use of these artifacts they must be aware of their existence. To capitalize on this opportunity we must devise new indexing methods that can integrate literature and data and capture their connectedness (RQ_1). We must also devise new algorithms for discovery that help researchers become aware of the existence of relevant materials while also recognizing the potential for new connections between them (RQ_2). Finally these new indexing methods and discovery algorithms must be compatible with existing systems to facilitate adoption (RQ_3).

2 Literature Review

By observing the current trends in cyberscholarship, digital library systems and open notebook science we can see the shape of a future scientific environment emerging. Features of this environment include the inclusion of research data alongside research publications and the availability of artifacts from all points along the scientific process. Pieces of this environment are being built opportunistically by researchers who are taking advantage of the tools on-hand while other pieces are being deliberately engineered with support from the National Science Foundation and other large funding agencies.

Tools for exploring scientific literature have been developed in the traditions of bibliometrics and literature related discovery. Separately analytical tools are being developed to address issues of information overload. Opportunities exist to combine techniques from these efforts to support development of tools for navigating within the broader scientific research information environment.

2.1 The Scientific Research Information Environment

2.1.1 Cyberscholarship, Cyberinfrastructure, Discovery and Innovation

Cyberscholarship refers to "new forms of research and scholarship that are qualitatively different from traditional ways of using academic publications and research data [AL07]." The use of the prefix "cyber" connotates digital and Internet-related qualities of these new forms. Often issues in cyberscholarship are positioned with issues in cyberinfrastructure. The infrastructure piece refers to software tools and hardware platforms that facilitate these new forms of research and scholarship. Discussions of the cyberinfrastructure often include free and open-source software systems running on supercomputers that are connected via the NSF TeraGrid[1]. NSF also has a program for "Cyber-Enabled Discovery and Innovation:"

> "Cyber-Enabled Discovery and Innovation (CDI) is NSF's bold five-year initiative to create *revolutionary* science and engineering research outcomes made possible by innovations and advances in computational thinking. Computational thinking is defined comprehensively to encompass computational concepts, methods, models, algorithms, and tools. Applied in challenging science and engineering research and education contexts, computational thinking promises a profound impact on the Nation's ability to generate and apply new knowledge. Collectively, CDI research outcomes are expected to produce paradigm shifts in our understanding of a wide range of science and engineering phenomena and socio-technical innovations that create new wealth and enhance the national quality of life [MRW09]."

The vision for the National Cyberinfrastructure was developed collaboratively and is articulated in [Nat07]. Execution of this vision is being managed by the National Science Foundation Office of Cyberinfrastructure[2]. Table 2 lists a number of recent NSF programs which are relevant to the research problem addressed in section 3. The work proposed in section 3 will complement these initiatives and contribute to the National Cyberinfrastructure.

[1] `http://www.teragrid`
[2] `http://www.nsf.gov/dir/index.jsp?org=OCI`

NSF Solicitation	Proposal Due Date
Software Development for Cyberinfrastructure [SBP+09]	February 26, 2010
Cyber-Enabled Discovery and Innovation [MRW08, MRW09]	February 4, 2010
CreativeIT [MTLS+09]	October 13, 2009[a]
Sustainable Digital Data Preservation and Access Network Partners (DataNet) [NS07]	November 13, 2008

[a]We have submitted a proposal for the CreativeIT solicitation which can be found in the supplementary materials.

Table 2: Selected NSF Solicitations

2.1.2 General Digital Library Systems

The library communities are seeing a trend in the shift from the dominance of physical local collections to digital federated collections of resources [Smi09]. Professional curation of digital collections requires much more technology and process than is directly supported by the popular file system tools available today. Extensive support for metadata is one discriminating characteristic separating digital library systems from simple file systems. Digital library systems provide a technology layer to support professional collections management on top of hierarchical file system and relational database management systems technologies. Two general purpose digital library systems are the Flexible Extensible Digital Object Repository Architecture[3] (FEDORA) and DSpace[4]. In May of 2009 the organizations that supported the development of FEDORA and DSpace merged to unify their efforts as the DuraSpace Organization [MKP09].

By taking a systems approach digital libraries generally include a server component for managing a collection. Container file formats can be considered a lightweight approach to metadata management. By utilizing a server process digital library systems impose a form of centralized control which is implicitly assigned to the organization running and managing the server process. Container file formats such as MPEG4[5] and HDF5[6] manage data by using a wrapper architecture. In general the data to be managed is encapsulated in a layer of metadata that describes the data. Multiple layers are enclosed in a single file. While the FEDORA and DSpace models also use multiple layers a differentiating characteristic of container file formats is that the metadata travels with the data rather than existing on a server as a reference. This allows for a decentralized approach to

[3]`http://www.fedora-commons.org`
[4]`http://www.dspace.org`
[5]`http://www.chiariglione.org/mpeg/standards/mpeg-4/mpeg-4.htm`
[6]`http://www.hdfgroup.org/HDF5/`

metadata management. Issues arising from incompatible data formats may consume ninety percent or more of the time spent on a data visualization project and increased usage of HDF5 or the newly proposed F5 container format are expected to help improve the situation [Ben09].

2.1.3 Domain Specific Digital Library Systems

While the DuraSpace projects seek to be as general as possible domain specific digital libraries have also emerged. "The Software Environment for the Advancement of Scholarly Research (SEASR), funded by the Andrew W. Mellow Foundation, provides a research and development environment capable of powering leading-edge digital humanities initiatives [SEA09]." SEASR through its partnership with the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign creates a bridge between humanities scholars and the supercomputing community. Supercomputing has historically focused on the needs of scientists working in the physical sciences. Despite this heritage management of scientific data remains an unsolved problem even in the physical sciences. The Science Commons organization seeks to make scientific research data more reusable and more useful. "If we can *systematically* increase our chances of making big discoveries and decrease the likelihood that we are ignoring information that we should be using then that's the best chance we have to get these breakthroughs in understanding about our bodies and about drugs [Dyl09, 1:29–1:44, emphasis added]."

2.1.4 Open Notebook Science

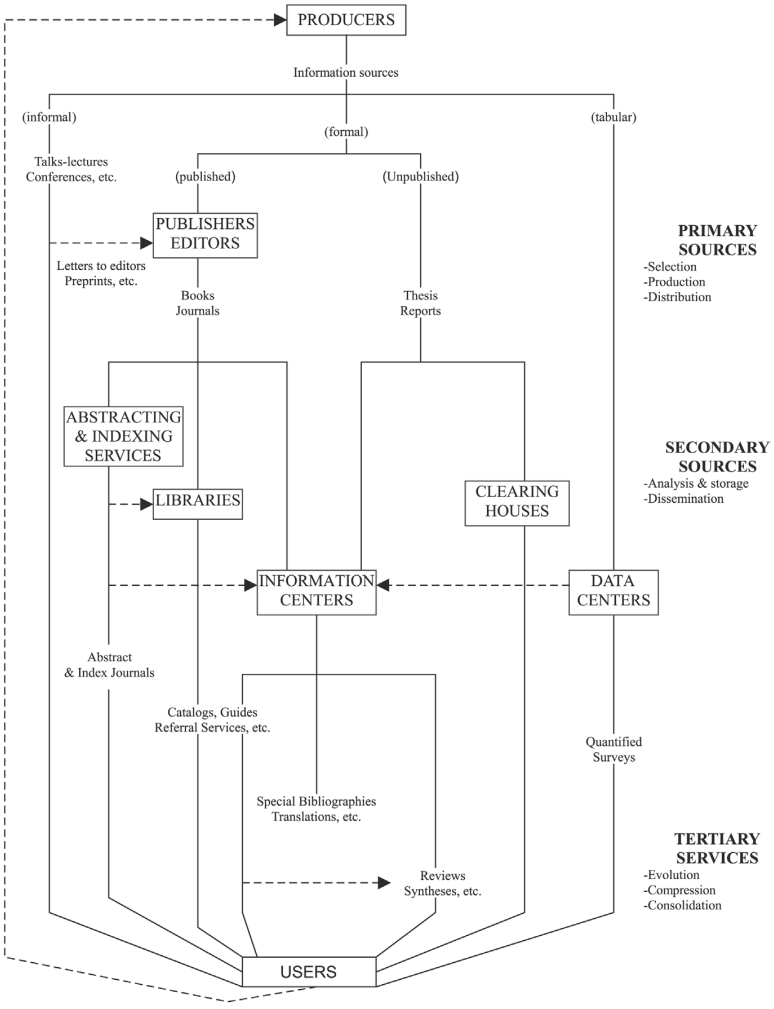
"Open Notebook Science is the practice of making the entire primary record of a research project publicly available online as it is recorded [Wik09]." Open Notebook Science shares the goal of publicizing research data that is advocated by the Science Commons. Open Notebook Science goes further and extends the objectives in the vein of the open source software development model to include both data and process. Exemplars of this approach include the UsefulChem[7] site by Jean-Claude Bradley and initiatives in India to support drug discovery [Sin08]. In many scientific institutions laboratory notebooks are kept as paper records by researchers. These notebooks are often archived in institutional libraries. A primary use of these records has been to support the patent process. While institutions have made attempts to digitize notebooks opening them to the

[7] `http://usefulchem.wikispaces.com`

public as they are being written represents a significant change in process.

2.1.5 Models of Scientific Communication

In 1971 the United Nations Educational, Scientific and Cultural Organization (UNESCO) and the International Council of Scientific Unions (ICSU) cooperated to publish the UNISIST model of scientific and technical communication shown in figure 1.



Note: Reproduced by permission of UNESCO

Source: UNISIST (1971, p. 26)

Figure 1: The original UNISIST model as reproduced in [SAH03].

In 2003 revisions and updates to this model were proposed that extend it to include Internet-based scholarly information [SAH03]. The extension of the 1971 model in 2003 revels that even the

most general model of scientific communication in 1971 was insufficient to completely classify the communications that emerged as a result of the introduction of Internet technologies. At the highest level the model identified three types of communications; primary sources, secondary sources and tertiary services as defined in Table 3. These three types of communication were superimposed on a flow model that described the artifacts during transitions and stages of information moving from producers to users. The extended model placed the Internet along side the entirety of the flow covering the full process from producer to user. It also added "Preprint Databases," "Scientific and Research Organizations Servers" and "Search Engines" as some of the significant new objects in the extended model [SAH03, Figure 5, p.303]. A final extension was to enclose the entire model within the boundary of a domain. This was done in recognition that "different epistemologies in a given domain will emphasise different knowledge sources [SAH03, p.305]." The revised model is shown is figure 2. Open Notebook Science as described in section 2.1.4 was not included in the 2003 revision however it might have occupied a space in the upper-right of figure 2 just above "Preprint Database."

2.1.6 Information Overload

The problem of too much information, or information overload, has become well recognized in popular culture as the issue pervades even personal information management such as email, tweets and Facebook updates [Zel09]. IDC predicts that "in 2011, the amount of digital information produced in the year should equal nearly 1,800 exabytes, or 10 times that produced in 2006. The compound annual growth rate between now [2008] and 2011 is expected to be almost 60% [GCM+08]." While the sheer volume of modern digital information creation is impressive the problem of managing large collections is perennial. One historical response to a sudden large increase in data volume has been to revisit indexing strategies as described in an account of managing intelligence data during World War II: "The indexes were not started as part of a great documentation plan, but simply emerged as response to the continuing and rapidly growing problems of controlling vast amounts of intelligence consequent on the successes in breaking Enigma and other encryption systems [Bru05]."

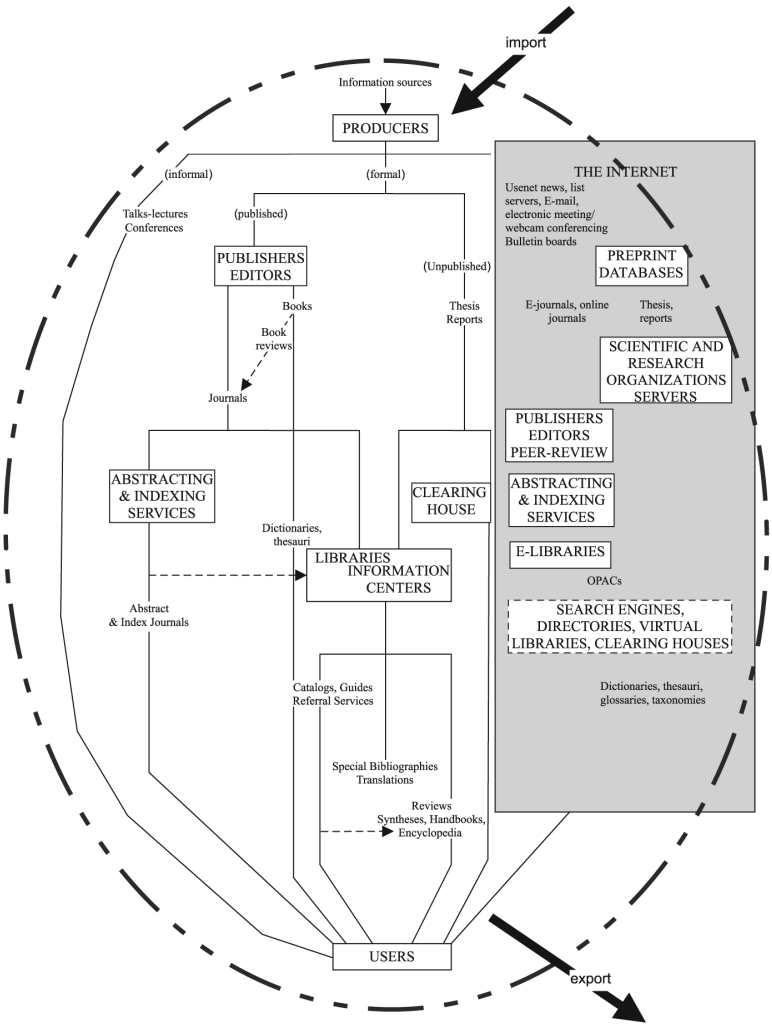


Figure 2: UNISIST model updated to reflect effects of the Internet [SAH03].

2.2 Navigating the Environment

2.2.1 Bibliometrics

Traditionally physical libraries have been primarily concerned with the management of their own local collections. This includes providing local cataloging and indexing services. With interlibrary loan programs the scope of resources exposed to a patron are expanded to include the collections of collaborating libraries. Access to such multi-institutional collections ranges in simplicity from searching each library's catalog individually, using federated search to automatically perform a search of each library's catalog, or use of a single catalog that contains aggregate data from all of the collections. Specialty indexes and manually authored domain specific bibliographies provide a

Primary Sources	"Primary literature is the researcher's and knowledge producer's primary medium for claiming original findings, theoretical analysis, empirical data etc.: Monographs... Journal articles... Critical-analysing reviews. Conference presentations. 'Grey' literature... Patents. Standards. [SAH03, p.318]" "Source literature is either literature produced in order to supply researchers with information (e.g. translation journals) or information produced to other purposes than research, but used as information by researchers (e.g. music and fiction)... *Data archives, Statistical documents, tabular documents* [SAH03, p.319, emphasis added]."
Secondary Sources	"Secondary literature / bibliographical literature. This is literature that registers, describes and organises the primary literature as well as the other categories (including the secondary literature itself). Secondary information systems are the core focus of the library, documentation, and information science profession. Bibliography is a discipline that studies this area: Subject bibliographies and bibliographical databases. ... [SAH03, p.319]."
Tertiary Services	"Tertiary literature / review literature / 'outlines.' This is literature summarising and synthesising knowledge in the primary literature: Handbooks... Review articles... *Data handbooks, tabular documents 2 (synthesising original statistical sources)* [SAH03, p.319–320]."

Table 3: High level classifications in the UNISIST model.

problem or domain view into the literature independent of aggregations by physical collection.

Eugene Garfield founded the Institute for Scientific Information (ISI) in 1960 which produced subject specific indexes of the literature [Gar09]. These indexes were unique in including the references cited by the articles included in the index. As these indexes became available electronically large scale analysis of the citation patterns became feasible. This fed the field of bibliometrics which led to the development of algorithms and visualization systems. Examples of modern systems include HistCite [GPI03], CiteSpace [Che06] and AuthorLink [LWB03]. Each of these systems provide a perspective on the literature that is algorithmically derived from the citation data as opposed to manual definition by expert bibliographers and index authors. In many cases the algorithmically identified perspectives invite users to discover novel relationships and new insights regarding the problem, domain or author being studied.

2.2.2 PageRank

The World Wide Web introduced a new multi-institutional document collection. This collection however lacks the professional curation and indexing practices followed by librarians. Search engines have attempted to expose the collection to users by generating their own indexes of the content and then providing a custom interface to the index. In 1998 Lawrence Page, co-founder of Google, filed a patent for the PageRank algorithm [Pag01]. The "field of the invention" is documented as:

> "This invention relates generally to techniques for analyzing linked databases. More particularly, it relates to methods for assigning ranks to nodes in a linked database, such as any database of documents containing citations, the world wide web or any other hypermedia database [Pag01]."

Page's contribution recognized that indexing the World Wide Web could be seen as an extension of citation analysis and bibliometrics by interpreting hyperlinks as bibliographic citations. Indeed Garfield's 1972 *Science* article [Gar72] is cited as a reference in the PageRank patent.

Within the field of bibliometrics the unit of analysis is limited to the bibliography of a work. More abstractly however the field has dealt with the connectedness of people and their ideas by using the measure of a citation as an indicator for behavioral phenomena regarding social networks, the formation of ideas over time and the current state of an intellectual domain. In a sense then the limits on bibliometric analysis are an artifact of the materials generally curated by traditional libraries and their subsequent indexing by ISI.

The trends identified in section 2.1 point to an expansion of the types of media collected and an increase in digital libraries. In particular data and supplementary materials from research works are increasing in availability. The unique contribution of ISI in 1960 to index a measure of connectedness through references cited can be compared to the the contribution of PageRank in 1998 to index the connectedness through hyperlinks of web pages. Each of these seminal contributions was heavily influenced by the nature of the media being indexed. In journal articles the citation serves as a behavioral indicator of intellectual constructionism. In web pages the hyperlink serves as an analogous behavioral indicator.

The metadata components of digital library systems and container file formats provide the opportunity to build networks of connected artifacts that transcend the explicit linkages established

by journal articles through citation and web pages through hyperlinking. The creation of such indexes is addressed by RQ_1 described in section 3.3.1.

2.2.3 Literature Related Discovery

Don Swanson pioneered the field of Literature Related Discovery in 1986 with the publication of "Undiscovered Public Knowledge" which opened:

> "Knowledge can be public, yet undiscovered, if independently created fragments are logically related but never retrieved, brought together, and interpreted. Information retrieval, although essential for assembling such fragments, is always problematic. The search process, like a scientific theory, can be criticized and improved, but can never be verified as capable of retrieving all information relevant to a problem or theory. This essential incompleteness of search and retrieval therefore makes possible, and plausible, the existence of undiscovered public knowledge... [Swa86b]."

The essential incompleteness of search and retrieval referred to by Swanson was explored in detail in 1968 by the philosopher Patrick Wilson with "Two kinds of power: an essay on bibliographical control [Wil68]." Although Wilson's work is not cited in "Undiscovered Public Knowledge" Swanson succeeds in integrating Wilson's ideas with Karl Popper's critique of positivism from the 1934 "Logik der Forschung" (The Logic of Scientific Discovery).

Swanson used computational analysis of citation data to infer syllogistic relationships between clusters of medical literature. It is notable that he used Garfield's ISI Science Citation Index via the DIALOG system for his work. With his approach he "...demonstrated that, at least qualitatively, the most successful attempts to treat Raynaud's syndrome tend to produce the same effects on certain blood parameters that dietary fish oil has been claimed to produce... [Swa86a]." This analytically discovered connection was then used as an initial hypothesis to be experimentally validated. It was later shown that fish oil did indeed alleviate the the symptoms of Raynaud's syndrome. For his work in Literature Related Discovery Swanson received the ASIS&T Award of Merit in 2000, the highest honor given by the American Society for Information Science and Technology [Swa01]. In his acceptance speech he remarked "Among all the people whose writing

have influenced and inspired me, an astonishingly high proportion of them have received an ASIS&T award, among them... Eugene Garfield... [Swa01]."

Work in Literature Related Discovery (LRD) has continued with Smalheiser [SS99] and Kostoff [Kos09] making notable contributions. The Arrowsmith [Swa08] system attempts to capture and expose much of the algorithmic work. The majority of the work in the field has continued to focus on the medical domain and by definition all of it continues to use the literature as the primary unit of analysis although some of the algorithms take advantage of the Medical Subject Headings (MeSH)[8] and domain ontologies to support natural language processing (NLP) aspects. I am not aware of any work that explicitly links LRD with heterogeneous data collections. Just as LRD has become possible with the digital indexing of citation data we can presume that digital indexes of heterogeneous data collections might facilitate new forms of discovery algorithms.

2.2.4 Data Related Discovery

The nomenclature of "Data Related Discovery" is not in common use. In the legal profession electronic discovery is a common term used to describe the process of electronically locating documents which are relevant to a particular case. I introduce the term here to refer to a particular subset of data mining and knowledge discovery and to contrast with Literature Related Discovery. As of November 22, 2009 a Google search for "Data Related Discovery" returns one hit and it is used analogously to electronic discovery on James Bowman's LinkedIn page[9].

The April 17, 2009 issue of *Science* included a pair of articles on computational support for scientific discovery that reported on techniques which narrowed the gap from the analysis of large volumes of data to the generation of scientific theory [WB09, SL09a, KRO+09]. Mass media coverage of these articles included headlines such as "Computer Program Self-Discovers Laws of Physics [Kei09]." In "Distilling Free-Form Natural Laws from Experimental Data [SL09a]" one of the experiments involved the input of motion tracking data recorded from a double-pendulum. "Without any prior knowledge about physics, kinematics, or geometry, the algorithm discovered Hamiltonians, Lagrangians, and other laws of geometric and momentum conservation [SL09a]." It is notable that this work was funded by the NSF CreativeIT program described in section 2.1.1.

[8] `http://www.nlm.nih.gov/mesh`
[9] `http://www.linkedin.com/in/jamesbowman`

An NSF press release reports that the algorithms were actually developed for work on self-repairing robots and then the researchers realized their general applicability to a large data space. Using the terminology from literature retrieval we can describe the algorithms used as search algorithms that covered the data space and produced minimally defined indexes to the data having maximal coverage of instances.

In "The Automation of Science" a robot "autonomously generated functional genomics hypotheses about the yeast *Saccharomyces cerevisiae* and experimentally tested these hypotheses by using laboratory automation [KRO+09]." Again an iterative search algorithm was used however with the novel contribution that the robot was able to affect the physical world and generate new data points to define the search space during the exploration iterations. Schmidt and Lipson close their article by describing the intended use of the work: "Scientists may use processes such as this to help focus on interesting phenomena more rapidly and to interpret their meaning [SL09a]."

Thus literature related discovery and data related discovery share commonalities in algorithmic design. Each use iterative data reduction and summarization to decompose a search space and each use pattern discovery to identify novel connections amongst elements of the decomposition. These classes of algorithms are generally explored in the fields of artificial intelligence and data mining [HH01].

While literature related discovery and data related discovery come from different historical traditions they also share a common use case. Each provide utility by helping a researcher focus in on interesting elements within a large collection. "Michael Atherton, a cognitive scientist who recently predicted that computer intelligence would not soon supplant human artistic and scientific insight, said that the program [Schmidt and Lipson] 'could be a great tool, in the same way visualization software is: It helps to generate perspectives that might not be intuitive' [Kei09]."

The obvious difference between LRD and Data Related Discovery is the unit of analysis. However the algorithms themselves are not tightly coupled to the raw input. Instead they operate on indexes or surrogates of the input, particularly as further iterations generate reductions and summarizations of the full information space. Therefore a combined information space of both literature and data has the potential to widen the scope of the discovery algorithms and therefore also increases the potential for finding connections across more widely disparate elements. The construction of such a combined information space is the subject of (RQ_1) described in section

3.3.1. The development of discovery algorithms to operate on such a space is the subject of (RQ_2) described in section 3.3.2. The systematic application of these new techniques is the subject of (RQ_3) described in 3.3.3.

During a teleconference describing their future work Schmidt and Lipson explain that a problem with their current technique is that although the algorithms find descriptive and succinct equations it is still a challenge to interpret the significance of those equations in the domain of study. They go on to say that this is a particularly difficult problem when analyzing bioinformatics data [SL09b]. A combined information space has the potential to address this problem. The high semantic density of a literature space can be used to contextualize patterns and unexplored elements of a data space when the two are correlated by a unified model.

2.2.5 Visual Analytics

In 2004 the United States Department of Homeland Security (DHS) chartered the National Visualization and Analytics Center (NVAC) at Pacific Northwest National Laboratory. Researchers from academia, industry and government collaborated to develop a five year research agenda and to define the grand challenges of the field. The results of this collaboration were published in "Illuminating the Path: The Research and Development Agenda for Visual Analytics [TC05]" one year later. The agenda was focused on addressing homeland security issues and intelligence analysis in particular. The grand challenges and research that the field produced are however generally applicable to any problems that require an understanding of complex data. Coincidentally the same year that DHS and NVAC were publishing a book on the future of intelligence analysis the American Society for Information Science and Technology was publishing a book on its history. In "Covert and Overt: Recollecting and Connecting Intelligence Service and Information Science" the editors report:

> "Originally, our intent was only to find some interesting speakers for a forthcoming professional conference. During the 2000 conference of the American Society for Information Science and Technology (ASIS&T), at a planning meeting for the Special Interest Group on History and Foundations of Information Science (SIG/HFIS), we undertook to arrange a session for the following year at which a panel of speakers would

> talk about their early backgrounds in intelligence work. It was already widely known, but rarely mentioned, that many of the people responsible for establishing the field of information science and for building ASIS&T into the leading professional association for the field had worked in intelligence agencies during World War II [WL05]."

The chapters of "Covert and Overt" are authored by individuals who provide personal accounts of their experience. Reading through them one can see that a pervasive theme is the problem of huge volumes of records being generated and the consequent challenge in developing and maintaining usable indexes. Many of the themes in the history of intelligence analysis and information science reappear in the Grand Challenge defined for Visual Analytics in "Illuminating the Path:"

> Grand Challenge: Enabling Profound Insights. One challenge underlies all of these objectives: the *analysis* of overwhelming amounts of disparate, conflicting, and dynamic information to identify and prevent emerging threats, protect our borders, and respond in the event of an attack or other disaster. This analysis process *requires human judgment* to make the best possible evaluation of incomplete, inconsistent, and potentially deceptive information in the face of rapidly changing situations [TC05, p.2].

While the problems of supporting human judgment with information are not new, modern increases both in volume (see section 2.1.6) and in kind (see section 2.1.5) of available information are notable. It is recognized that visual representations of information can take advantage of aspects of human cognition in powerful ways. This has long been known in the field of cartography [Mac95]. Leveraging these cognitive capacities of the human visual system for information management with modern interactive computer graphics is therefore a promising path forward.

Lee S. Strickland, former intelligence officer at the Central Intelligence Agency and professor at the University of Maryland College of Information Studies, explicitly identifies the need for tools to integrate literature and data:

> "Another key is addressing the volume of information–a veritable tsunami–and the need for tools. In short, the totality of information far exceeds the ability of any organization to effectively and completely analyze and render judgments. And there are several aspects to this issue. One is that textual information must be captured

and must be retrievable. Another is that the textual information or structured data quickly outstrips the working capability of the mind to retain and thus analyze. *Yet another is the necessity to integrate that unstructured text information with structured data.* These issues present a critical requirement: analytical software (tools) to work on the problems of entity and relationship extraction from texts as well as the analysis of the resulting data (e.g., the discovery of trends or links that are quite simply not obvious to the human analyst)[Str05, p. 164, emphasis added]."

The Visual Analytics Science and Technology (VAST) Contests and Challenges were created to help support the development of new visual analytics tools by providing datasets with a hidden ground truth [PFG08, WCH+06]. The use of a shared dataset helps to facilitate comparative evaluations of new tools and designs. The 2006 and 2007 VAST contest datasets made use of corpus of textual data and award winning teams generally leveraged entity extraction algorithms in combination with interactive visualizations [SGLS07, SGZ08, GLP+07]. In 2008 the format of the competition was changed with the introduction of mini-challenges [GPL+08]. With this change the mini-challenges were generally composed of structured data while the volume of textual data was reduced. The inclusion of image data remained small. Success in the Grand Challenge in 2008 required integrating multiple structured data sources and contextualizing the data by the narrative found in the textual elements. Winning entries generally made use of a graph data structure to integrate the heterogeneous sources while each structured data source was also given its own customized interactive visualization to support exploration [PPR+08, PSSM08, CTP+08].

2.2.6 Scenario Visualization

In "Scenario Visualization: An Evolutionary Account of Creative Problem Solving" Robert Arp contends that the human visual system has specifically evolved to allow humans to perform nonrountine creative problem solving by making novel connections between previously unrelated information [Arp08a].

> "Unlike routine problem solving–which deals with associative connections within familiar perspectives–nonroutine creative problem solving entails an innovative ability to make connections between wholly unrelated perspectives or ideas [Arp08a, p. 9]."

By viewing the human visual system as a hierarchical, modular system defined by information filtering and flow Arp shows how humans chunk surrogates for raw images and then perform transformation operations on those surrogates to build novel connections. Arp defines scenario visualization as "a conscious activity whereby visual images are selected, integrated, and then transformed and projected into visual scenarios for the purpose of solving problems in the environments one inhabits [Arp08a, p. 2]."

The essence of the discovery algorithms can be interpreted as a subset of the general process described by Arp. In the context of LRD the chunks are coded as clusters of document surrogates (metadata) and the purpose of the algorithm is to identify connections between unrelated chunks that may be relevant to solving a problem in the world.

3 Problem Statement

One specific mechanism of scientific discovery is the creative process of making novel connections between previously disconnected bodies of knowledge [FMC07, Swa86b]. By observing the current trends in cyberscholarship, digital library systems and open notebook science we can begin to see a future scientific environment emerging. In this environment primary, secondary and their derivative source materials will be publicly available in digital collections accessible over the Internet. Collections will include artifacts from all steps in the scientific process rather than only the narrative of the final results from journal articles. These developments constitute an expansion of the fundamental composition of the scientific research information environment. Visions of this future environment purport qualitative differences in the research process marked by systematic support for discovery. The work proposed here is designed to improve our understanding of this new scientific research information environment. The understanding will be developed sufficiently to inform the creation of tools and processes that systematically support the process of making novel connections between previously disconnected bodies of knowledge within this new environment. With the accomplishment of this work researchers will be able to use scientific data more effectively for the creation of new knowledge.

3.1 Problem being addressed

Researchers faced with the problem of finding relevant data and other artifacts that are not part of the literature are lacking tools to support their search. Keyword search is ineffective for artifacts that are not composed of descriptive vocabulary, such as datasets. Indexes that are organized by topicality or metadata provide limited access points to such artifacts. This problem is particularly acute when the artifacts have a general applicability or future utility that was not predicted by the metadata curator. Indexes based on behavioral indicators of relevance such as citation records or hyperlinks are not feasible for artifacts which do not utilize these conventions.

3.2 Motivation

Repeating experiments to collect data that has already been collected by someone else wastes time, money and equipment. Often valuable research data was not reused because it was not available. Recent trends are resulting in the availability of much more data. Unfortunately it may still go unused because it cannot be found or its relevance is not recognized at the time it is needed. Research into the nature of how relevance is defined and recognized in collections of data is necessary.

3.3 Research Questions

This section enumerates three research questions to be addressed through experimentation in the dissertation work. For each question a specific aim is defined as well as a number of candidate solutions. It is expected that further elaboration of the candidate solutions will lead to specific methodologies for experimentation.

RQ_1	What are the generic mechanisms by which connections can be established between a corpus of literature and a data repository?
RQ_2	Given a corpus of literature which has been connected to a data repository which algorithms are most effective for supporting the discovery of novel connections among the two?
RQ_3	How can digital library technologies be extended to support the discovery of novel connections among collections of literature and data?

Table 4: Research Questions

3.3.1 Establishing connections

Research Question (RQ_1) What are the generic mechanisms by which connections can be established between a corpus of literature and a data repository?

Specific Aim ($SA_{1.1}$) Develop a set of general methods for connecting a corpus of literature with a related data repository.

Candidate Solution ($CS_{1.1}$) Apply natural language processing against literature to systematically identify identifiers in the controlled vocabulary of a domain that reference data in a repository. Solutions in this set will take the form of pattern analysis algorithms that are first tuned to a particular domain, such as NCBI GI or Accession numbers appearing in Bioinformatics publications. Domain specific dependencies will be assessed for their generalizability to other domains. Professor Tony Hu's laboratory has expertise working in this solution space.

Candidate Solution ($CS_{1.2}$) Using the supplementary materials URLs in literature scrape the resources and ingest them into a customized data repository. Since this technique implicitly establishes a link between the data and the literature the next challenge will be to curate the collection of supplementary materials so that connections might be established within the data repository itself. One approach to investigate is the use of a Fedora Commons models to facilitate automatic ingestion and MIME type assignment for finding connections between objects.

Preliminary Work Participation in the VAST 2008 Challenge provided an opportunity to model a collection of heterogeneous data. Our full entry is archived at the National Institute of Standards and Technology [PCM+08] and a short paper describing the entry was presented at the IEEE VisWeek 2008 conference [PPR+08].

> "The VAST 2008 Challenge scenario concerned a fictitious, controversial socio-political movement. Participants were provided with an excerpt from the movement's manifesto and the following four data sets, one for each mini-challenge:
>
> - cell phone records over a 10 day period

- a chronicle of migrant boat journeys with passenger lists, launch and landing sites and landing/interdiction status
- a catalog of wiki edits to a page discussing the movement
- geospatial data of an evacuation from a building in which a bomb exploded [GPL+08]."

Individual team members focused on specific mini-challenges and built customized interactive visualization systems to analyze each data set. Although a single visualization design tool was used for each mini-challenge it remained difficult to generate an integrated view. A minimalistic model required customized ingest functions to handle parsing the separately formatted data files. The ingest functions had to be programmed manually. The instantiation of five matrices of the data is shown in figure 3 with references to the parsing routines, "LoadCellCallsNumeric()," "LoadParsedWikiEditsPage()," "LoadMigrantData()," "LoadOccupantsRFIDAssignments()," "LoadOccupantsRFIDPathways()." The generalizability of this approach is thus limited by the manual effort needed to program customized parsers for any given structured dataset that is to be included. Digital library systems have the potential to overcome this limitation by exposing a common Application Programming Interface (API) to multiple data streams.

With all of the data loaded into a common data structure it was then transformed into an associative network to facilitate processing by graph theoretic algorithms. This step provided an opportunity to annotate the structure with hypotheses that had been developed during the course of individual team member's work on the mini-challenges. Figure 4 shows the encoding used to add the hypotheses to the model.

In traditional data models the raw input is usually kept separate and distinct from hypotheses, assumptions and conclusions made concerning the data. Often the raw input is stored in data files, databases and statistical tools while the hypotheses, assumptions and conclusions are reserved for the narrative of textual reports. A notable exception to this is the Analysis of Competing Hypotheses model that explicitly captures relations between hypotheses and their supporting evidence [Heu99]. Interpreted in the context of the UNISIST model described in section 2.1.5 we can observe that the delineation of content as a primary, secondary or tertiary source often affects the form of management applied to its collection and archival. The pervasiveness of this relationship is often implicit and might be considered a historical artifact of librarianship grounded in a scientific pos-

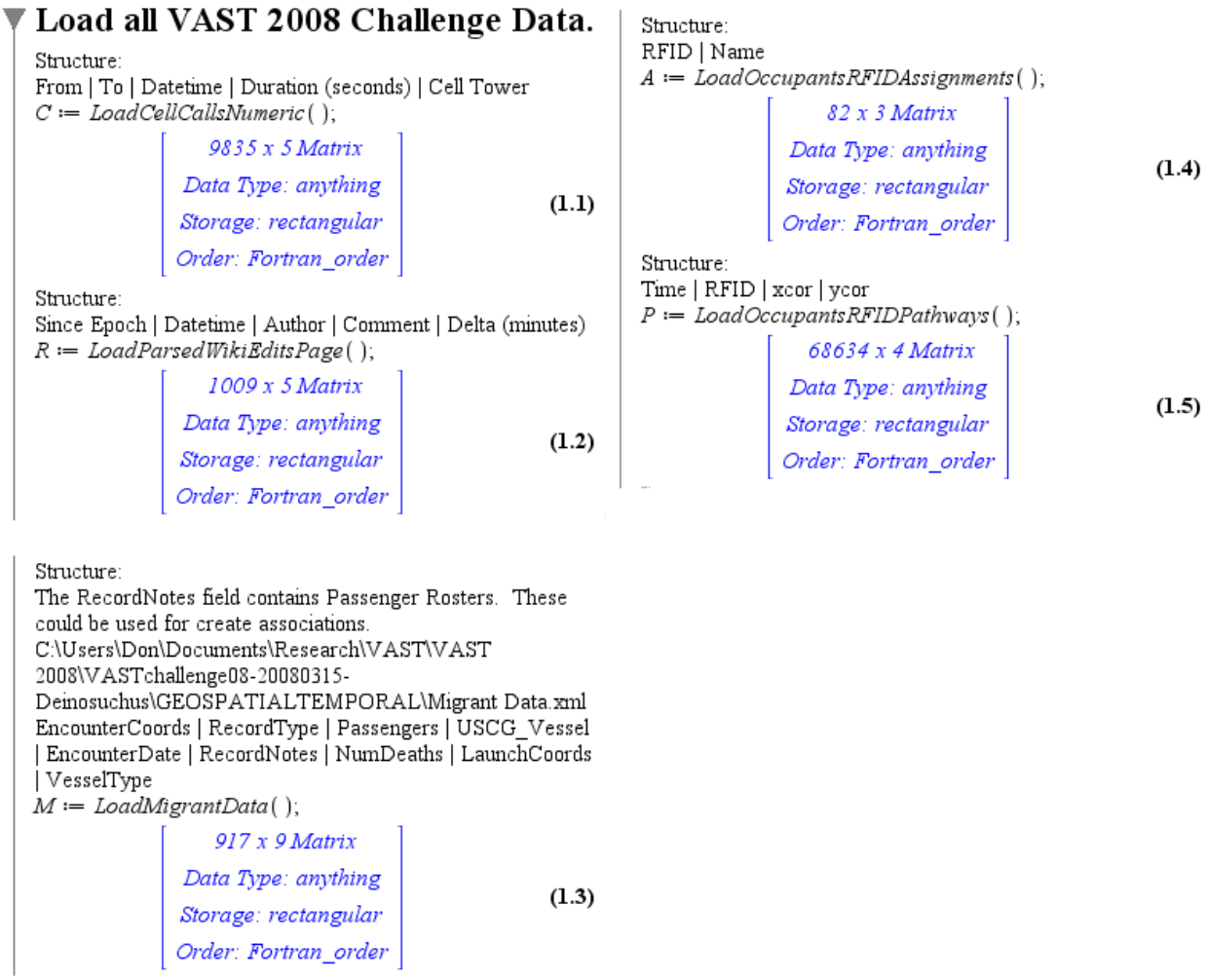


Figure 3: All VAST 2008 Challenge data unified into matrices [PCM+08, Figure 1].

itivism. The separate collections operate as stovepipes reinforcing the separation of these sources rather than their connectedness.

Our VAST 2008 model used simplistic single character prefixes (R = RFID Tag, G = Given Name, S = Surname, I = Coast Guard Intercept Manifest, H = Hypothesis, T = Telephone) to capture both the type and provenance of the elements in the model. Container file formats and digital library systems have much more robust metadata support to allow for type and provenance encoding at large scales.

```
# Hypotheses
H_casualties := vertex;
fprintf(fd, "%d \"H Casualties\"\n", vertex);
vertex := vertex + 1 :

H_suspects := vertex :
fprintf(fd, "%d \"H Suspects\"\n", vertex) :
vertex := vertex + 1 :
```

```
# Results of the Evacuation Mini-Challenge.
for i from 1 to 82 do
 # Add the Casualties Hypothesis.
 if evalb(A[i, 1]in {18, 19, 56, 36, 76, 50, 39, 78, 65, 60, 47, 69}) then
  fprintf(fd, "%d %d\n", AID[i] + 1, H_casualties);
  fprintf(fd, "%d %d\n", H_casualties, AID[i] + 1);
 end if;

 # Add the Suspects Hypothesis.
 if evalb(A[i, 1]in {21, 1, 29, 44, 56}) then
  fprintf(fd, "%d %d\n", AID[i] + 1, H_suspects);
  fprintf(fd, "%d %d\n", H_suspects, AID[i] + 1);
 end if;
end do:
```

Figure 4: Adding hypotheses directly to the integrated model [PCM+08, Figure 7].

3.3.2 Algorithms for discovering novel connections

Research Question (RQ_2) Given a corpus of literature which has been connected to a data repository which algorithms are most effective for supporting the discovery of novel connections among the two?

Specific Aim ($SA_{2.1}$) Develop a set of pattern discovery algorithms that can identify novel candidate connections that are meaningful in the domain.

Candidate Solution ($CS_{2.1}$) Represent the relations among the literature and data as a graph. Use SUBDUE [CH07b, CH07a, HCD94] to identify common sub-graphs as candidate structures in the domain. Search for connections that would create additional instances of such structures. Contextualize the candidate instances with domain knowledge and evaluate their semantics and significance in the domain. Techniques for predicting missing links in networks are also described in [CMN08].

Candidate Solution ($CS_{2.2}$) Combining visual analytics techniques with dimensionality reduction algorithms create maps of the information space that can be interactively navigated by users for discovery and exploration tasks. Algorithms for projection of large graphs such as [ADWM04, FT07, New04, CNM04] will likely be useful in this task.

Mind's Eye Algorithms The nomenclature of "Mind's Eye Algorithms" is not in common use. A Google search performed on November 23, 2009 returned one hit with an instance of the three words occurring separated by a comma and not used as a single compound word. I introduce the term here to contrast with neural network algorithms and to refer to a model of human creative connection building based on biological models of the human visual system.

Models of the human brain have inspired the class of neural network algorithms in the field of artificial intelligence. The algorithms are modeled on our understanding of the interaction of neurons in the brain [Nil98]. Just as models of biological neurons and synapses have inspried this class of learning algorithms models of the human visual system described in [Arp08b] in the context of non-routine creative problem solving and in [Mac95] in the context of data reduction and symbolism can be used for the design of a new class of algorithms. These Mind's Eye Algorithms will use multiple simultaneous surrogate representations to represent an information space and patterns of linkages to identify novel and meaningful connections within the space. An example of the affects of localized brain damage illustrates the loose coupling between identifier metadata and outline representations of three dimensional figures in the brain:

> "Studies of patients whose association areas are damaged have shown that different representations of an object are stored differently. For example, a person suffering from *associative visual agnosia* whose posterior parietal cortex has been damaged can identify objects by drawing them but cannot name them. Conversely, a person suffering from *apperceptive visual agnosia* whose occipital lobes are damaged can name objects but cannot identify them to draw them (Farah, 199). Another example is *prosopagnosia*, the inability to recognize familiar faces or learn new faces that results from damage to the IT cortex. People who suffer from prosopagnosia, although unable to process or recall faces, still can process and recall other objects, such as animals and tools (Geschwind, 1979). The studies indicate that the visual image is a product of multiple representations in the brain, each having their own neural correlates and each concerned with a different aspect of the visual image. This implies that there is no *one* deposit memory storage area. Rather, there are multiple storage areas, and recollection is itself a process of building up disparate pieces of information [Arp08c, p. 80]."

A historical tradition in library science holds that there is a correctness to the indexing of materials. The index should be consistent, repeatable and predictable. Once an artifact is indexed its metadata is generally held static even while new materials are introduced. Models of creative problem solving in the human visual system however indicate that biological indexing is almost the antithesis of this ideal. Human creativity can be interpreted as the process of constantly rebuilding the indexes so as to create new connections and relationships between existing information. This process appears to be most fully stimulated when a new piece of information is introduced that does not fit easily into the existing index structures. The idea of misfit has been explored in cognitive science [Sim96] and it has been related to the phenomena of having an insight in visual analytics [Nor06, YKSJ08, CZGR09].

The development of Mind's Eye Algorithms is a specific, deliberate and tangible strategy for the creation of systematic methods to support human creative problem solving. Many of the architectural elements necessary for implementation are already available. A loose coupling between surrogate representations of literature or data and their related metadata can be encoded in a graph data structure linking container file formats or objects encoded with the FEDORA model. Dynamic graph instantiation and weighting can be performed relative to a loosely defined problem. User interest can be expressed by allowing a user to identify a number of objects and then requesting an expression of the types of relations than can be found among those objects. Such a query interface is analogous to current query by example techniques. Research in data mining on interestingness provides an entry point for exploration of the core components of the algorithms [HH01]. Substructure discovery in graphs provides another class of algorithms that may have utility for this purpose [HCD94, PC08].

Mind's Eye Algorithms extend LRD algorithms through the use of loose coupling that eliminates the dependency on citation data. They also extend Data Related Discovery algorithms by accepting heterogeneous data sources. They overcome limitations of data mining and pattern discovery algorithms by facilitating a natural domain contextualization of the discovered patterns.

Preliminary Work Continuing the example from section 3.3.1 we see that the input to the algorithmic discovery phase is the heterogeneous associative network established by techniques developed in pursuit of (RQ_1). Figure 5 shows the associative network created for the VAST 2008

Challenge.

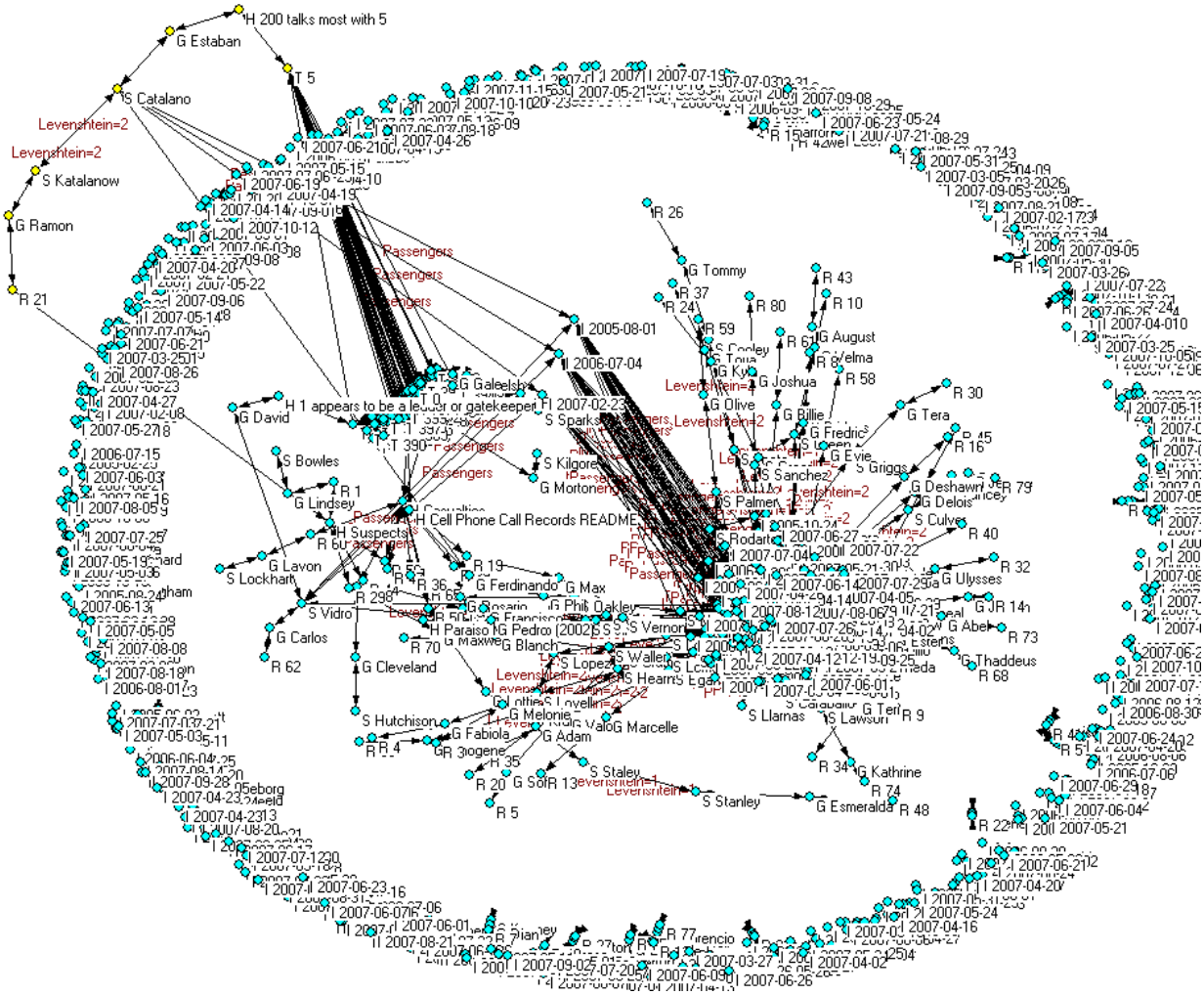


Figure 5: Associative network of heterogeneous data [PCM$^+$08, Figure 8].

The storyline of the challenge indicates a relationship between the Vidro and Catalano families. The nature of this relationship however is unknown. Graph analytics on the integrated model support the exploration of the relationship shown in figure 6. Levenshtein distance measures are computationally added to the model. In this instance the semantics added computationally allow the user to recognize the significance of a misspelling in the data that was missed during exploratory analysis. Additionally a relationship between two RFID tags in the hospital motion data is given a closer look although initial analysis of the motion data did not indicate a connection between these two points. Thus by using a simple search criteria a complex pattern is discovered that

is contextualized by the provenance of specific supporting data. Using the pattern as an initial hypothesis analysts were able to initiate an exploration of data that was otherwise thought to be unrelated. In the context of the synthetic story it was found that these two individuals made attempts in 2005 and 2006 at illegal immigration into the United States which were intercepted by the Coast Guard. Then in 2007 they successfully landed in Mexico. They later were both present at the hospital during the bombing. Individual analysis of the Coast Guard data and of the hospital RFID data had missed this connection. Collaboration between the analysts working on these two sets also failed to identify the pattern. This use of the model for search and retrieval given explicit search criteria is one general use case.

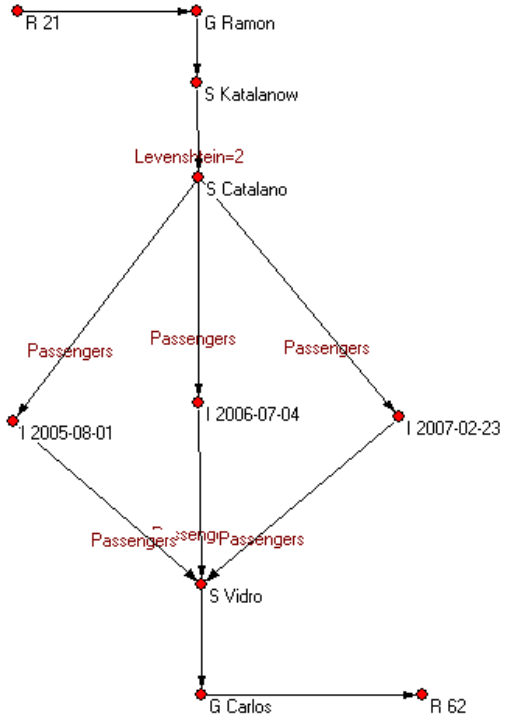


Figure 6: An interesting subgraph [PCM$^+$08, Figure 10].

Another use case for the model is the automatic discovery of patterns that are inherently interesting. Operating in a mixed-initiative mode the discovery algorithms might run as background agents constantly examining the data for deviations from user established or domain specific rule-sets. In this mode it is not necessary for the user to initiate a search for relevant material. Figure 7 shows the discovery of a connection that violates domain logic. RFID tag 56 is hypothesized to

be both a suspect and a casualty. Computational discovery of this relation triggers action by the human team of analysts to reassess their assumptions. A team from HRL Laboratories, LLC used a variation on this technique for the 2009 VAST Challenge [ALH09]. After building the network model they encoded domain information into the pattern discovery process. They then relied on anomaly detection to identify hypotheses rather than relying on human analysts to creatively formulate the starting points.

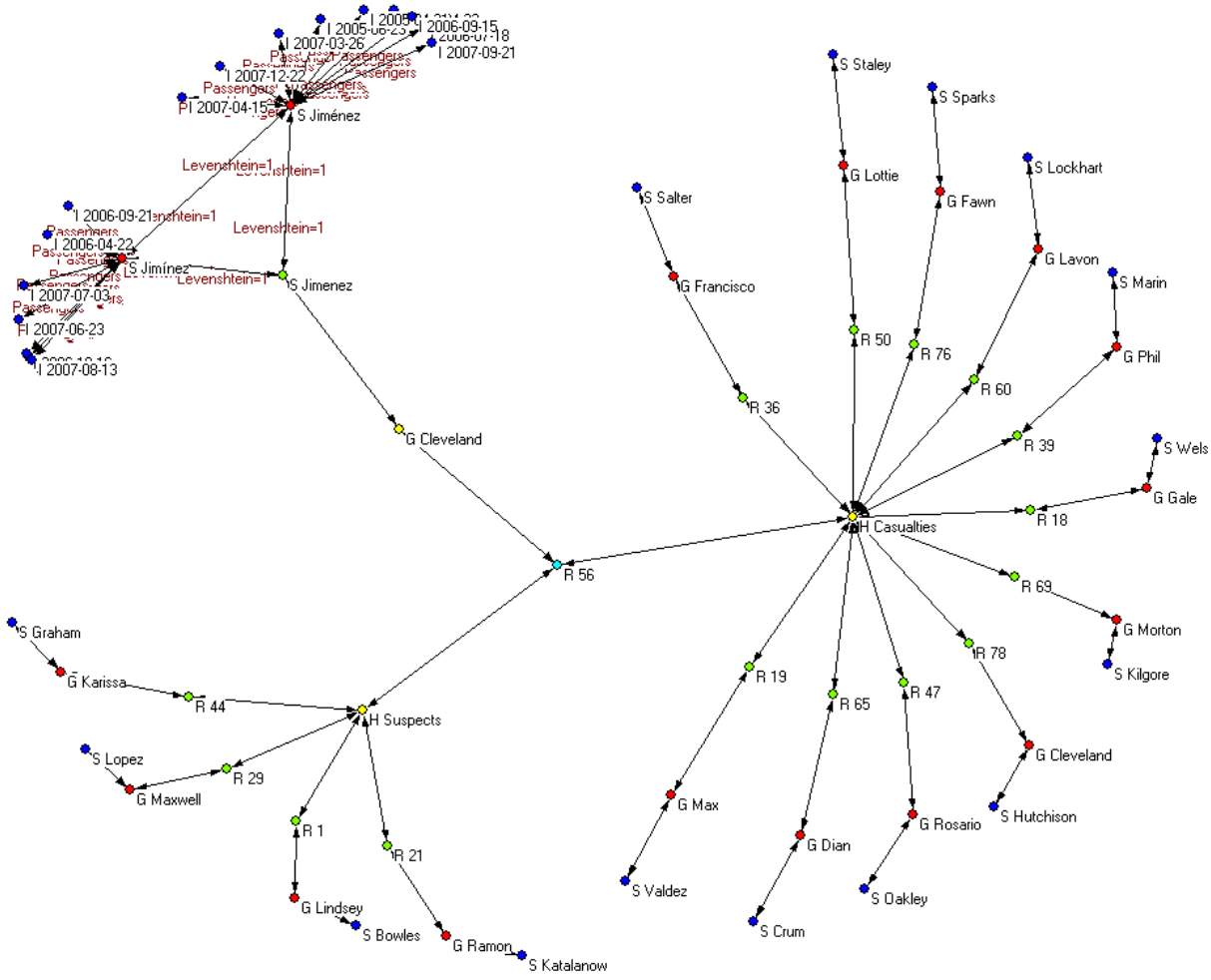


Figure 7: Violation of domain constraints [PCM+08, Figure 11].

Although the preceding examples are from the domain of intelligence analysis the combination of similar associative network models and pattern discovery algorithms has the potential to support the creative process of making novel connections in many domains, including data driven scientific

research. Additionally the preceding examples did not leverage the scientific literature or make use of traditions in bibliometrics. The newly initiated "Influenza Sequence Mapping Project[10]" can be used to support experimentation applying these techniques to influenza research. The "UsefulChem Project[11]" exposes a collection of heterogeneous data that is specifically designed to be broadly useful. Automatic identification of connections between UsefulChem content and other projects might serve as novel hypotheses for analytical verification.

3.3.3 Extending Digital Libraries

Research Question (RQ_3) How can digital library technologies be extended to support the discovery of novel connections among collections of literature and data?

Specific Aim ($SA_{3.1}$) Identify leverage points in digital library technologies that can be extended to support the algorithms developed for ($SA_{2.1}$) and the methods developed for ($SA_{1.1}$) leading to systematic application of these methods. This will constitute a contribution to practice.

Candidate Solution ($CS_{3.1}$) Digital library technologies such as DuraSpace, DSpace and Fedora Commons are open-source projects that include extension mechanisms. Components of these technologies will be used to expand the generalizability of the of the specific aims. This will allow the methods developed for discovery of novel connections to be incorporated with existing work-flows that are supported by digital library technologies.

Candidate Solution ($CS_{3.2}$) Digital library technologies commonly support traditional keyword search and retrieval techniques using form based methods. The discovery of novel connections represents a new form of access to data that more directly supports the creative process of generating new ideas rather than the pursuit of existing questions. Visualization interfaces and mixed initiative approaches as described in [HKR+08] may be more appropriate designs for exposing discovered candidate connections.

[10] `http://cluster.ischool.drexel.edu/~st96wym4/flumap/`
[11] `http://usefulchem.wikispaces.com/`

4 Evaluation Strategies

A challenging aspect of evaluation in an information systems project is achieving ecological validity. The context of use is often a critical factor determining system utility and acceptance in the field. The living laboratory [Chi09] approach is one of the most effective means for achieving high ecological validity. With a living laboratory a live system is provided to users who then use it for their own purposes within their own work processes and work environment. Google's Beta approach is a good example of a living laboratory. Google is able to release new features to a sub-population of their application users who then use the features in the course of their normal routine. By comparing the performance of different features exposed to different sub-populations Google can evaluate them with high ecological validity.

Yang, Cohen and Hersh discriminate between extrinsic and intrinsic evaluation approaches in selecting an evaluation strategy for their Gene Information Clustering and Summarization System [YCH09]. Extrinsic evaluation "measures how useful the system is to the intended end users" while intrinsic evaluation "measures how good the system is [YCH09]." They found it useful and practical to develop evaluation tasks to be performed by Bioinformatics researchers, their target audience, which allowed the researchers to use the their own microarray data that they were analyzing at the time of the evaluation. This approach represents a hybrid or limited living laboratory since the users were able to focus on their own data although the tasks they performed were specified by the evaluators.

Swanson's approach to evaluation was in the style of the existence proof. Rather than performing an detailed user evaluation of the system he simply used his technique personally to identify candidate hypotheses that he then published in the medical literature. The utility of the system was measured in terms of the number of useful hypotheses it helped to generate.

The purpose of visualization systems is to generate insight [Nor06, YKSJ08, CZGR09]. Unfortunately defining insight is a challenge and measuring it even more challenging. Similarly with supporting the creative process of making novel connections it is a challenge to define a novel connection and then to measure how many have been made. Evaluation methods for visual analytics are addressed in "Illuminating the Path [LP05]." Information retrieval has used metrics of precision and recall for evaluation. Precision is a measure of the relevance of a given result set while recall

is a measure of how many relevant documents were missed by the search.

The VAST Challenge uses two evaluation strategies [PGS+08]. For the open contest a ground truth is deliberately hidden across the data sets. Quantitative measures are calculated based on the number of ground truth elements found in a solution. Qualitative measures are assessed based on expert opinion of the quality of the approach. Winning teams are then invited to a closed session during the IEEE VisWeek conference. During this session the winning systems are given to professional intelligence analysts to use on a new problem set. A team of observers then take detailed notes on the use of the system in this session. Qualitative content analysis of the notes, video and audio of the session then provide a means to assess the use of the system. With multiple teams and systems comparative assessments can also be made.

It has also been proposed that distributed cognition is a useful framework for evaluating information visualization systems [LNS08]. Using distributed cognition the system boundary is expanded to included the users of the systems and their interactions. Interactions mediated by the systems as well as external interactions may be studied.

Brandes and Lerner's analysis of Wikipedia behavior is an example of the use of quantitative log data analyzed and interpreted to allow observation of group dynamics [BL08]. Using only the revision history of Wikipedia pages they are able to identify controversial issues, the size and composition of factions and even make statements about individual behavior. By analyzing log data they are able to achieve features of the living laboratory such as examining the behavior of people using the systems for their own purposes in their own context of use. The key enabler for the success of their techniques was to identify quantitative elements of the data that could be used as indicators of user behavior. By establishing a threshold on the time duration between revisions it could be inferred that a user was responding to a revision made by another user or simply adding new content. Observing response revision patterns and distributions allowed for the detection of edit wars, controversial issues which were the subject of those wars and factions of members who held similar or opposing viewpoints. This style of analysis is consistent with the tradition in bibliometrics of interpreting a citations as a behavioral indicator of relevance. Had the live Wikipedia system been available a full living laboratory approach might be possible, whereby engineering changes to Wikipedia could be paired with observation of user behavior.

An ideal evaluation strategy for the candidate solutions proposed in section 3 would be to iden-

tify behavioral indicators that could be calculated based on system usage. Log files from a digital library system or custom visualization interface could be used for this purpose. This would allow for statistical profiles of connection building to be observed. Tools for discovering novel connections could then be introduced and their effect on user behavior observed and interpreted. The primary challenge of this approach would be to integrate with a running system with a sufficient user base. This approach is also described as using computational social science to design information retrieval algorithms [Pel09].

The VAST Challenge evaluation strategies are also feasible approaches to evaluation of the candidate solutions proposed in section 3. Using these strategies a scientific dataset containing a known discovery could be presented to users. Variations on tool designs would be the dependent variable for experimental evaluation. Metrics indicating the closeness of a participant to a discovery set could be used to measure system effectiveness.

References

[ADWM04] A. T. Adai, S. V. Date, S. Wieland, and E. M. Marcotte, "LGL: Creating a Map of Protein Function with an Algorithm for Visualizing Very Large Biological Networks," *Journal of Molecular Biology*, vol. 340, no. 1, pp. 179–190, June 25 2004. [Online]. Available: http://dx.doi.org/10.1016/j.jmb.2004.04.047

[AL07] W. Y. Arms and R. L. Larsen, "The future of scholarly communication: Building the infrastructure for cyberscholarship," National Science Foundation and the Joint Information Systems Committee, Report of a workshop held in Phoenix, Arizona April 17 to 19, 2007, September 26 2007. [Online]. Available: http://www.sis.pitt.edu/~repwkshop/NSF-JISC-report.pdf

[ALH09] D. Allen, T.-C. Lu, and D. Huber, "Detecting and analyzing relationships among anomalies," *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pp. 255–256, October 12–13 2009. [Online]. Available: http://dx.doi.org/10.1109/VAST.2009.5334426

[Arp08a] R. Arp, *Scenario visualization : an evolutionary account of creative problem solving.* Cambridge, Mass.: MIT Press, 2008.

[Arp08b] ——, "Scenario visualization, creative problem solving, and evolutionary psychology," in *Scenario visualization : an evolutionary account of creative problem solving.* Cambridge, Mass.: MIT Press, 2008, ch. 5, pp. 133 – 166.

[Arp08c] ——, "The visual system," in *Scenario visualization : an evolutionary account of creative problem solving.* Cambridge, Mass.: MIT Press, 2008, ch. 3, pp. 57 – 89.

[Ben09] W. Benger, "On safari in the file format jungle–why can't you visualize my data?" *Computing in Science and Engineering*, vol. 11, no. 6, pp. 98–102, November/December 2009. [Online]. Available: http://dx.doi.org/10.1109/MCSE.2009.202

[BL08] U. Brandes and J. Lerner, "Visual analysis of controversy in user-generated encyclopedias," *Information Visualization*, vol. 7, no. 1, pp. 34–48, 2008. [Online]. Available: http://dx.doi.org/10.1057/palgrave.ivs.9500171

[Bru05] R. Brunt, "Some aspects of indexing in british intelligence, 1939-1945," in *Covert and Overt: Recollecting and Connecting Intelligence Service and Information Science.* Medford, NJ: Information Today, Inc., 2005, pp. 81 – 106. [Online]. Available: http://books.infotoday.com/asist/CovertOvert.shtml

[CH07a] D. J. Cook and L. Holder, "Subdue - graph based knowledge discovery," Pullman, WA, 2007. [Online]. Available: http://www.subdue.org

[CH07b] ——, "Subdue manual version 1.4," January 11, 2007 2007, software Manual. [Online]. Available: http://www.subdue.org

[Che06] C. Chen, "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 359–377, 2006. [Online]. Available: http://dx.doi.org/10.1002/asi.20317

[Chi09] E. H. Chi, "A position paper on 'living laboratories': Rethinking ecological designs and experimentation in human-computer interaction," *Lecture Notes in Computer Science*, vol. 5610, pp. 597–605, 2009. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-02574-7_67

[CMN08] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, May 2008. [Online]. Available: http://dx.doi.org/10.1038/nature06830

[CNM04] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, p. 066111, 2004. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.70.066111

[CTP+08] L. Chien, A. Tat, P. Proulx, A. Khamisa, and W. Wright, "Grand challenge award 2008: Support for diverse analytic techniques - nspace2 and geotime visual analytics," in *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*. Columbus, OH: IEEE, October 19-24 2008, pp. 199–200.

[CZGR09] R. Chang, C. Ziemkiewicz, T. M. Green, and W. Ribarsky, "Defining insight for visual analytics," *IEEE Computer Graphics and Applications*, vol. 29, no. 2, pp. 14–17, March/April 2009. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/MCG.2009.22

[Dyl09] J. Dylan. (2009, November 5) Science commons. Video. Science Commons. [Online]. Available: http://sciencecommons.org/about/science-commons-dylan-video/

[FMC07] L. Fleming, S. Mingo, and D. Chen, "Collaborative brokerage, generative creativity, and creative success," *Administrative Science Quarterly*, vol. 52, no. 3, pp. 443–475, September 2007. [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=27501417&site=ehost-live

[FT07] Y. Frishman and A. Tal, "Multi-Level Graph Layout on the GPU," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1310–1319, Nov.-Dec. 2007. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2007.70580

[Gar72] E. Garfield, "Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies," *Science*, vol. 178, no. 4060, pp. 449–526, November 3 1972. [Online]. Available: http://dx.doi.org/10.1126/science.178.4060.471

[Gar09] ——. (2009, November 18) Eugene garfield, ph.d. career overview. University of Pennsylvania. [Online]. Available: http://www.garfield.library.upenn.edu/overvu.html

[GCM+08] J. F. Gantz, C. Chute, A. Manfrediz, S. Minton, D. Reinsel, W. Schlichting, and A. Toncheva, "The diverse and exploding digital universe: An updated forecast of worldwide information growth through 2011," IDC, An IDC White Paper - sponsored by EMC, March 2008. [Online]. Available: http://www.emc.com/digital_universe

[GLP+07] C. Görg, Z. Liu, N. Parekh, K. Singhal, and J. Stasko, "Jigsaw meets blue iguanodon - the vast 2007 contest," in *IEEE VAST '07*, Sacramento, CA, October 2007, pp. 235–236. [Online]. Available: http://www.cc.gatech.edu/~stasko/papers/vast07-contest.pdfhttp://www-static.cc.gatech.edu/gvu/ii/jigsaw/

[GPI03] E. Garfield, A. I. Pudovkin, and V. S. Istomin, "Why do we need algorithmic historiography?" *Journal of the American Society for Information Science and Technology*, vol. 54, no. 5, pp. 400–412, 2003. [Online]. Available: http://dx.doi.org/10.1002/asi.10226

[GPL+08] G. Grinstein, C. Plaisant, S. Laskowski, T. O'Connell, J. Scholtz, and M. Whiting, "Vast 2008 challenge: Introducing mini-challenges," in *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*. Columbus, OH: IEEE, October 19-24 2008, pp. 195–196.

[HCD94] L. Holder, D. J. Cook, and S. Djoko, "Substructure discovery in the subdue system," in *AAAI Workshop on Knowledge Discovery in Databases*, 1994. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.44.6602

[Heu99] R. J. J. Heuer, *Psychology of Intelligence Analysis*. Central Intelligence Agency, 1999. [Online]. Available: https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/psychology-of-intelligence-analysis/index.html

[HH01] R. J. Hilderman and H. J. Hamilton, *Knowledge Discovery and Measures of Interest*. Kluwer Academic Publishers, 2001.

[HKR+08] C. G. Healey, S. Kocherlakota, V. Rao, R. Mehta, and R. St. Amant, "Visual perception and mixed-initiative interaction for assisted visualization design," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 2, pp. 396–411, March / April 2008. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2007.70436

[Kei09] B. Keim, "Computer program self-discovers laws of physics," *Wired Science*, April 2 2009. [Online]. Available: http://www.wired.com/wiredscience/2009/04/newtonai/

[Kos09] R. N. Kostoff, "A systematic approach to alternative medical procedures," *BioScience*, vol. 59, no. 9, pp. 734–735, October 2009, viewpoint article. [Online]. Available: http://dx.doi.org/10.1525/bio.2009.59.9.2

[KRO+09] R. D. King, J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova, A. Sparkes, K. E. Whelan, and A. Clare, "The automation of science," *Science*, vol. 324, no. 5923, pp. 85–89, April 3 2009.

[LNS08] Z. Liu, N. J. Nersessian, and J. T. Stasko, "Distributed cognition as a theoretical framework for information visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1173–1180, November/December 2008. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TVCG.2008.121

[LP05] S. Laskowski and C. Plaisant, "Evaluation methodologies for visual analytics," in *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, J. J. Thomas and K. A. Cook, Eds. Los Alamitos, CA: IEEE Computer Society, 2005, pp. 150–157. [Online]. Available: http://nvac.pnl.gov/agenda.stm#book

[LWB03] X. Lin, H. D. White, and J. Buzydlowski, "Real-time author co-citation mapping for online searching," *Information Processing & Management*, vol. 39, no. 5, pp. 689–706, Sep 2003. [Online]. Available: http://dx.doi.org/10.1016/S0306-4573(02)00037-7

[Mac95] A. M. MacEachren, *How Maps Work: Representation, Visualization, and Design*. New York, NY: The Guilford Press, 1995. [Online]. Available: http://www.guilford.com/cgi-bin/cartscript.cgi?page=pr/maceachren.htm&sec=summary&dir=geo/tech&cart_id=459591.18039

[MKP09] C. M. Morris, M. Kimpton, and S. Payette, "Fedora commons and dspace foundation join together to create duraspace^{TM} organization," May 12 2009, press Release. [Online]. Available: http://duraspace.org/pressrelease.php

[MRW08] E. Misawa, T. Russell, and K. Whang, "Cyber-enabled discovery and innovation (cdi)," Program Synposis, National Science Foundation, Arlington, Virginia, Program Solicitation NSF 08-604, November 5 2008. [Online]. Available: http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503163

[MRW09] ——, "Cyber-enabled discovery and innovation (cdi)," National Science Foundation, Arlington, VA, Program Solicitation NSF 10-506, November 5 2009. [Online]. Available: http://www.nsf.gov/publications/pub_summ.jsp?WT.z_pims_id=503163&ods_key=nsf10506

[MTLS+09] M. L. Maher, B. K. Tuller, A. La Salle, J. Peckham, C. L. Bloebaum, A. M. de Strulle, J. I. Lane, A. T. Desena, and E. Arkilic, "Creativeit," National Science Foundation, Program Solicitation NSF 09-572, 2009. [Online]. Available: http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=501096

[Nat07] National Science Foundation Cyberinfrastructure Council, "Cyberinfrastructure vision for 21st century discovery," National Science Foundation, Arlington, VA, Tech. Rep. NSF 07-28, March 2007. [Online]. Available: http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf0728

[New04] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, p. 066133, 2004. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.69.066133

[Nil98] N. J. Nilsson, "Neural networks," in *Artificial Intelligence: A New Synthesis*. San Francisco, CA: Morgan Kaufmann Publishers, 1998, ch. 3, pp. 37–57.

[Nor06] C. North, "Toward measuring visualization insight," *IEEE Computer Graphics and Applications*, vol. 26, no. 3, pp. 6–9, May/June 2006. [Online]. Available: http://dx.doi.org/10.1109/MCG.2006.70

[NS07] L. Nowell and S. Spengler, "Sustainable digital data preservation and access network partners (datanet)," National Science Foundation, Arlington, VA, Program Solicitation NSF 07-601, September 28 2007. [Online]. Available: http://www.nsf.gov/publications/pub_summ.jsp?WT.z_pims_id=503141&ods_key=nsf07601

[Pag01] L. Page, "Method for node ranking in a linked database," United States of America Patent 6,285,999, September 4, 2001. [Online]. Available: http://patft.uspto.gov/netacgi/nph-Parser?Sect2=PTO1&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&d=PALL&RefSrch=yes&Query=PN%2F6285999

[PC08] D. Pellegrino and C. Chen, "Automatic hypothesis generation and evaluation by network structure content analysis and visualization," in *Annual U.S. Department of Homeland Security University Network Summit*, Washington, DC, March 18 2008. [Online]. Available: http://www.orau.gov/DHSSummit/2008/materials.htm

[PCM+08] D. Pellegrino, C. Chen, A. MacEachren, P. Mitra, C.-C. Pan, A. Robinson, M. Stryker, and C. Weaver, "North-east visualization and analytics center (nevac) team entry," in *VAST Challenge Portal*. National Institute of Standards and Technology, 2008. [Online]. Available: http://vac.nist.gov/challenge2008.html

[Pel09] D. Pellegrino, "Overcoming groupthink in distributed heterogeneous data analysis," April 23 2009, poster presented at Drexel University Research Day, Philadelphia, PA. [Online]. Available: http://www.donpellegrino.com/is/Pellegrino-ResearchDay2009.pdf

[PFG08] C. Plaisant, J.-D. Fekete, and G. Grinstein, "Promoting insight-based evaluation of visualizations: From contest to benchmark repository," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 1, pp. 120–134, January/February 2008. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2007.70412

[PGS+08] C. Plaisant, G. Grinstein, J. Scholtz, M. Whiting, T. O'Connell, S. Laskowski, L. Chien, A. Tat, W. Wright, C. Görg, Z. Liu, N. Parekh, K. Singhal, and J. Stasko, "Evaluating visual analytics at the 2007 vast symposium contest," *IEEE Computer Graphics and Applications*, vol. 28, no. 2, pp. 12–21, March/April 2008.

[PPR+08] D. Pellegrino, C.-C. Pan, A. Robinson, M. Stryker, J. Luo, C. Weaver, P. Mitra, C. Chen, I. Turton, and A. MacEachren, "Grand Challenge Award: Data Integration - Visualization and Collaboration in the VAST 2008 Challenge," in *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*. Columbus, OH: IEEE, October 19-24 2008, pp. 197–198. [Online]. Available: http://dx.doi.org/10.1109/VAST.2008.4677384

[PSSM08] J. Payne, J. Solomon, R. Sankar, and B. McGrew, "Grand challenge award: Interactive visual analytics palantir: The future of analysis," in *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*. Columbus, OH: IEEE, October 19-24 2008, pp. 201–202.

[SAH03] T. F. Søndergaard, J. Andersen, and B. Hjørland, "Documents and the communication of scientific and scholarly intformation: Revising and updating the unisist model," *Journal of Documentation*, vol. 59, no. 3, pp. 278–320, 2003. [Online]. Available: http://dx.doi.org/10.1108/00220410310472509

[SBP+09] J. Schopf, A. Blatecky, M. Parashar, P. Bogden, and M. McClure, "Software development for cyberinfrastructure (sdci)," National Science Foundation, Arlington, VA, Program Solicitation NSF 10-508, November 19 2009. [Online]. Available: http://www.nsf.gov/publications/pub_summ.jsp?WT.z_pims_id=5174&ods_key=nsf10508

[SEA09] SEASR. (2009, November 5) About seasr. Website. [Online]. Available: http://seasr.org/

[SGLS07] J. Stasko, C. Görg, Z. Liu, and K. Singhal, "Jigsaw: Supporting investigative analysis through interactive visualization," in *IEEE VAST '07*, Sacramento, CA, October 2007, pp. 131–138. [Online]. Available: http://www.cc.gatech.edu/~stasko/papers/vast07-jigsaw.pdfhttp://gvu.cc.gatech.edu/ii/jigsaw

[SGZ08] J. Stasko, C. Gorg, and L. Zhicheng, "Jigsaw: supporting investigative analysis through interactive visualization," *Information Visualization*, vol. 7, no. 2, pp. 118–132, Summer 2008.

[Sim96] H. A. Simon, *The Sciences of the Artificial*, 3rd ed. Cambridge, MA: The MIT Press, 1996.

[Sin08] S. Singh, "India takes an open source approach to drug discovery," *Cell*, vol. 133, no. 2, pp. 201–203, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/B6WSN-4S9G1SD-3/1/6a13f210185df623e1ccdb040f482f7f

[SL09a] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," *Science*, vol. 324, no. 5923, pp. 81–85, April 3 2009. [Online]. Available: http://dx.doi.org/10.1126/science.1165893

[SL09b] ——. (2009, April 2) Maybe robots dream of electric sheep, but can they do science? - teleconference. Teleconference Audio. National Science Foundation. Arlington, VA. [Online]. Available: http://nsfgov.http.internapcdn.net/nsfgov_vitalstream_com/podcast/lipson_schmidt.mp3

[Smi09] M. Smith, "A research agenda for an academic research library, perspectives from mit," October 29 2009, presentation at Drexel University.

[SS99] D. R. Swanson and N. R. Smalheiser, "Implicit text linkages between medline records: Using arrowsmith as an aid to scientific discovery," *Library Trends*, vol. 48, no. 1, pp. 48–59, Sum 1999. [Online]. Available: http://find.galegroup.com.ezproxy2.library.drexel.edu/gtx/infomark.do?contentSet=IAC-Documents&docType=IAC&type=retrieve&tabID=T002&prodId=AONE&docId=A57046526&userGroupName=drexel_main&version=1.0&searchType=CCLSearchForm&source=gale&infoPage=infoMarkPage

[Str05] L. S. Strickland, "Knowledge transfer: Information science shapes intelligence in the cold war era," in *Covert and Overt: Recollecting and Connecting Intelligence Service and Information Science*, R. V. Williams and B.-A. Lipetz, Eds. Medford, NJ: Information Today Inc., 2005, ch. 12, pp. 147–166. [Online]. Available: http://books.infotoday.com/asist/CovertOvert.shtml

[Swa86a] D. R. Swanson, "Fish oil, raynauds syndrome, and undiscovered public knowledge," *Perspectives in Biology and Medicine*, vol. 30, no. 1, pp. 7–18, Fal 1986. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/3797213?itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_RVDocSum&ordinalpos=1

[Swa86b] ——, "Undiscovered public knowledge," *Library Quarterly*, vol. 56, no. 2, pp. 103–118, Apr 1986. [Online]. Available: http://dx.doi.org/10.1086/601720

[Swa01] D. Swanson, "Asist award of merit acceptance speech: On the fragmentation of knowledge, the connection explosion, and assembling other people's ideas," *Bulletin*

of the American Society for Information Science and Technology, vol. 27, no. 3, pp. 12–14, 2001. [Online]. Available: http://dx.doi.org/10.1002/bult.196

[Swa08] ——, "Welcome to arrowsmith 3.0," Chicago, IL, p. The home page for the Arrowsmith system., 2008. [Online]. Available: http://d-swanson.uchicago.edu/

[TC05] J. J. Thomas and K. A. Cook, "Illuminating the path: The research and development agenda for visual analytics," Los Alamitos, CA, 2005. [Online]. Available: http://nvac.pnl.gov/agenda.stm#book

[WB09] D. Waltz and B. G. Buchanan, "Automating science," *Science*, vol. 324, no. 5923, pp. 43–44, April 3 2009. [Online]. Available: http://dx.doi.org/10.1126/science.1172781

[WCH+06] M. A. Whiting, W. Cowley, J. Haack, D. Love, S. Tratz, C. Varley, and K. Wiessner, "Threat stream data generator: creating the known unknowns for test and evaluation of visual analytics tools," in *BELIV '06: Proceedings of the 2006 AVI workshop on BEyond time and errors*. Venice, Italy: ACM, 2006, pp. 1–3.

[WFE08] S. Wiedenback, M. Fazio, and M. Elliott, "The PhD Degree Program Description and Procedures," 2008, drexel University College of Information Science and Technology, 2008 Edition. [Online]. Available: http://www.ischool.drexel.edu/content/documents/PhDProgramDescription.pdf

[Wik09] Wikipedia contributors, "Open notebook science," November 5 2009. [Online]. Available: http://en.wikipedia.org/w/index.php?title=Open_Notebook_Science&oldid=321776139

[Wil68] P. Wilson, *Two kinds of power: an essay on bibliographical control*. Berkeley, University of California Press, 1968.

[WL05] R. V. Williams and B.-A. Lipetz, "Covert and overt: Recollecting and connecting intelligence service and information science," Medford, NJ, p. 250, 2005. [Online]. Available: http://books.infotoday.com/asist/CovertOvert.shtml

[YCH09] J. Yang, A. Cohen, and W. Hersh, "Evaluation of a gene information summarization system by users during the analysis process of microarray datasets," *BMC Bioinformatics*, vol. 10, p. (Suppl 2):S5, February 5 2009. [Online]. Available: http://www.biomedcentral.com/1471-2105/10/S2/S5

[YKSJ08] J. S. Yi, Y.-a. Kang, J. T. Stasko, and J. A. Jacko, "Understanding and characterizing insights: How do people gain insights using information visualization?" in *BELIV '08: Proceedings of the 2008 conference on BEyond time and errors*. Florence, Italy: ACM, April 5 2008, pp. 1–6.

[Zel09] N. Zeldes, "How to beat information overload," *IEEE Spectrum*, October 2009. [Online]. Available: http://spectrum.ieee.org/computing/it/how-to-beat-information-overload/0