Katy Börner. (2006) Semantic Association Networks: Using Semantic Web Technology to
Improve Scholarly Knowledge and Expertise Management.
In Vladimir Geroimenko & Chaomei Chen (Eds.) Visualizing the Semantic Web,
Springer Verlag, 2nd Edition, Chapter 11, pp. 183-198.

Chapter 11

Semantic Association Networks: Using Semantic Web Technology to Improve Scholarly Knowledge and Expertise Management

Katy Börner

School of Library and Information Science, Indiana University, Bloomington, IN, 47405, USA, katy@indiana.edu

11.1 Introduction

This chapter introduces *Semantic Association Networks* (SANs), a novel means of using semantic web technology to tag and interlink scientific datasets, services (e.g., algorithms, techniques, or approaches), publications (e.g., papers, patents, grants), and expertise (i.e., author and user information) to improve scholarly knowledge and expertise management. Among other ends, the proposed SANs

- Facilitate new types of searches, e.g., the retrieval of all authors that worked with dataset *x* or all papers that used algorithm *y*;
- Ease the reuse of datasets and services, thus increasing the reproducibility of results;
- Enable dataset/algorithm/result comparisons at the data/code/implication level;
- Exploit data access and data origin logs to indicate the usefulness of resources and the reputation of authors.

Section 11.2 outlines the need for improved scholarly knowledge and expertise management by discussing major trends in the evolution of science and our current means to access scholarly knowledge and expertise. Section 11.3 introduces *Semantic Association Networks*, and their usage for improving storage, correlation, analysis, access, and management of scientific knowledge and expertise. Section 11.4 discusses opportunities and socio-technical challenges for the implementation of SANs. Section 11.5 contains concluding remarks and presents a vision for the future of scholarly publishing knowledge and expertise management.

The content and style of this chapter were motivated by my research on knowledge domain visualizations (KDVis) (Börner et al., 2003; Shiffrin & Börner, 2004). KDVis apply advanced data mining and information visualization techniques to analyze, correlate, and visualize the semantic space of researchers, publications, funding, etc. The resulting visualizations can be utilized to objectively identify major research areas, experts, institutions, grants, publications, journals, etc., in a research area of interest. In addition, they can assist in the identification of interconnections, of import and export of research between fields, of the dynamics (speed of growth, diversification) of scientific fields, scientific and social networks, and the impact of strategic and applied research funding programs. Hence, KDVis give researchers and practitioners a global view of the structure and evolution of a research area.

High quality datasets that report scholarly activity are required to map science on a large scale and in a comprehensive manner. Consequently, I spent a considerable amount of my time gaining access to high quality datasets including publications, patents, and grants. All datasets are stored in a multi-terabyte Oracle database to facilitate the cross-correlation and search of the diverse datasets, e.g., interlinking author and investigator names from diverse publication and grant datasets, linking papers/patents to cited papers/patents, etc. While creating this database, I was amazed to learn how little care we take of our collective scholarly knowledge. While some communities support excellent efforts – such as the digitization of all *Physical Review* papers going back to 1893 – many institutions and organizations do not have the funds or have not prioritized the digitization and preservation of datasets for general or expanded use. Vast quantities of valuable, irreplaceable, scholarly digital datasets have already been lost forever due to the implementation of new data formats or the migration of computer systems.

However, access to high-quality, cross-referenced digital datasets covering decades and centuries of scholarly work, reliably preserved for future generations, are the basis for better means of scholarly data access, management, and ultimately the understanding and fostering of scientific progress. We now have the technological means to digitize, store, and make available scholarly data and to research science using the scientific methods of science as suggested by Derek J. deSolla Price exactly 40 years ago (1965). It is my hope that this chapter inspires others to work towards the implementation of better means to improve access to scholarly knowledge and expertise and to facilitate the analysis and communication of the structure and evolution of science.

11.2 Scientific Trends and Current Means to Access Knowledge and Expertise

Recently, a number of papers have been published that express a growing dissatisfaction with the existing communication system in terms of what can be published, by what means, for what price, with which latency, and using what copyright (Lynch, 2003; Henry, 2003; Smith et al., 2003; Williams et al., 2003) (Van de Sompel et al., 2004). The papers propose to rethink the scholarly communication system in terms of the size and media types of the 'published unit', open access, copyright issues, but they also discuss value added services that improve access and management.

Herbert Van de Sompel and his colleagues discuss a scholarly communication system that scholars deserve (2004). They provide a detailed overview of the changing nature of scholarly research and the resulting new demands on scholarly communication systems. They argue that most of today's research is highly collaborative, network-based and data-intensive and that fundamental changes in scholarly communication are needed. They point out the need for a system that, among others, offers interoperability across publishing venues, persistent identifiers to publications (and I will later on argue also to authors, services and datasets), and supports navigation across publishing venues, e.g., between journal papers from different publishers but also patents or grants. A scholarly 'interoperability substrate' is needed to design value-adding service that are not 'vertically' locked, see Fig. 1.

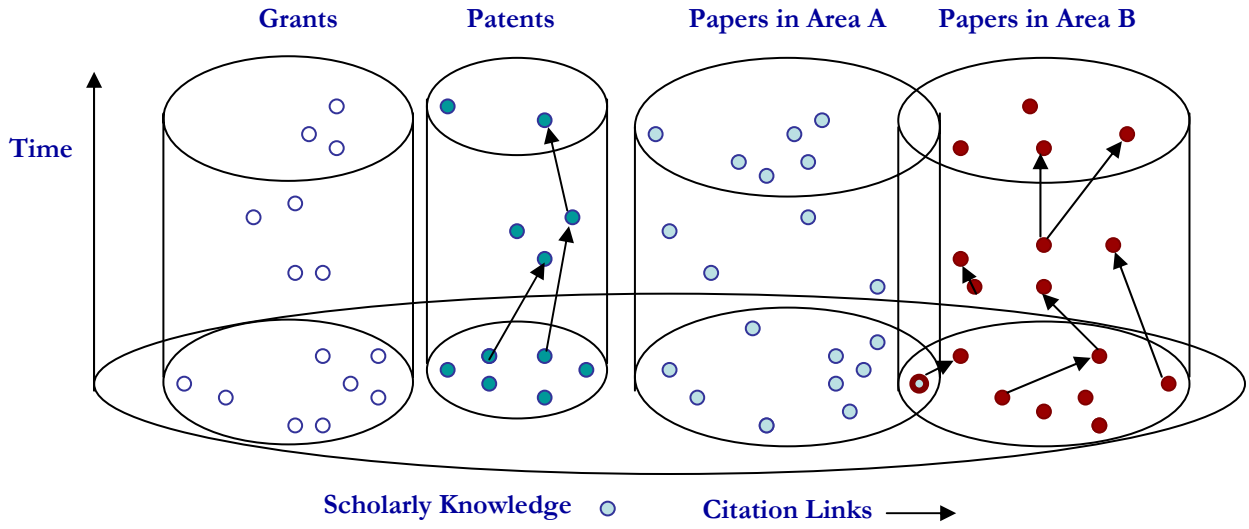

Figure 1: The interoperability and cross linkage problem. Many but not all of today's scholarly datasets, e.g., papers, patents, grants, are stored and made available so that 'vertical' citation linkages can be traversed. There are very few instances in which datasets of different origin and/or type are 'horizontally' interlinked.

Subsequently, I review general trends in the amount and quality of data produced and used by today's scientists, currently available means to store scholarly data, and the knowledge and expertise management that needs to be supported.

Many, if not all, areas of science are experiencing a 'tidal wave' of data.[1] In some fields of science, the volume of data doubles every 18 months (Scholtz, 2000). Data intensive sciences such as astronomy, geography, biology, and physics generate data at a great rate. Large digital sky surveys are becoming dominant, generating 10-100 terabytes - and soon petabytes - per survey (Djorgovski, 2004). Theoretical simulations are also becoming more complex and can generate terabytes of data. The generated datasets exhibit orders of magnitude larger, more complex, and heterogeneous than in the past.

The advent of the computer, the birth of the WWW, and the widespread availability of digitized information have vastly changed our ways of accessing information (Borgman, 2000). As Richard Shiffrin puts it: "The traditional method involved books, reference works and physical materials on library shelves, most of which had been verified for accuracy by one or another authority. Now, we sit at computers and cast our net into a sea of information, much of which is inaccurate or misleading."(Shiffrin, 2004). Today, a major challenge is to help people navigate the growing, partially interconnected 'knowledge webs'. People need guidance in the selection of the most correct or relevant data source, support in the examination of search results, and a reliable estimate of the quality of retrieved results. While the (scientific) promise of having a major part of humanity's knowledge at our fingertips is tremendous, the usefulness of this knowledge depends critically on our ability to extract relevant information and to make sense of it.

In this tidal wave of data, scholars act as information sources and sinks; they need to gain access to high quality information and they are interested in effortlessly diffusing their own scientific results. An infrastructure that aims to support collective scholarly knowledge and expertise management will need to take advantage of the perceptual and cognitive abilities of their human users. It should be technology enabled, but driven by scientific needs. The infrastructure will need to federate existing massive, complex datasets (measures, simulated data, publications), content (data, metadata) services, standards for data representation, and the identification of analysis/compute services. It should be openly distributed on a large scale so that all scholars can participate, and also dynamically evolving so that future needs can be met.

Over the past decades, diverse systems and approaches have been proposed that support the storage, interlinkage, and linkage-based retrieval of facts. Here I report major historical milestones in chronological order.

The *Shepard's Citations* published for and used by the legal profession since 1873 interlink legal texts, thus making linkage-based search possible. In 1938, Herbert G. Wells described his conception of the future information center in his book *World Brain*, inspiring numerous efforts to create a global repository of interlinked knowledge (1938). Vannevar Bush (1945), in his seminal work, *As we may think*, developed a new concept called 'memex' for the automatic storage and efficient access of books, records, and individual communications. Eugene Garfield's work on the *Science Citation Index* was deeply inspired by the *Shepard's Citations*. He envisioned and later implemented the citation indexes for science as an 'association-of-ideas index' (Garfield, 1955a). Influenced by V. Bush's ideas, Douglas Engelbart (1963) describes one of the first hypertext systems. Also in 1963, Theodor H. Nelson coined the words *'hypertext'* and *'hypermedia'* to describe his vision of worldwide hypertext – a universe of interactive literary and artistic works and personal writings deeply intertwined via hyperlinks. In 1960, Theodor H.

[1] http://www.arl.org/forum04/djorgovski.html

Nelson founded *Xanadu*[2], a project devoted to the design of a system that supports two-way, unbreakable links, deep version management, incremental publishing, document-document intercomparison, copyright simplification and softening, and origin connection (e.g., by retrieving quoted contents from the virtual original of the author or rights holder such that exact royalty payment for each download can be calculated). In 1988, Thinking Machines Inc. developed the *Wide Area Information Server* (WAIS) a distributed system to search index databases on remote computers. In 1989, Tim Berners-Lee architected the *World Wide Web*, an internet-based hypermedia initiative for global information sharing. Even with its one-way, fragile links and no inherent management of version or contents, the Web – billions of web pages authored and interlinked by millions of people around our planet – is the largest 'knowledge web' in existence today.

For the last decade, our main means of accessing the knowledge stored in digital libraries, repositories, or the Web has been the search engine. Search engine usage resembles 'charging a needle' with a search query and 'sticking it' into a 'haystack' of unknown size and consistency. Upon 'pulling the needle out', one checks the linearly sorted items that got 'stuck' on the needle. However, one typically does not get any information on the appropriateness of the 'haystack' probed, the algorithm applied to retrieve the items, or the quality of the data retrieved. Some systems support linkage-based search (e.g., finding all papers that cite paper *x* or all web pages that link to page *y*), see Fig. 2, but no search engine provides a more global view of the data searched or retrieved.

The dominant usage of search engines in combination with the accelerating pace of information generation and nearly constant human cognitive abilities leads to a general loss of global knowledge on the content of knowledge repositories, and on the general structure and evolution of scientific disciplines. No one today has a global view of mankind's knowledge. In fact, our bird's eye views are at best one meter above the landscape of science whereas a global view would only be possible from a 100 meter height.[3] Sadly, our distance above ground is decreasing steadily as the amount of information is growing.

The difference between the search for concrete facts and gaining a global overview of a research domain is illustrated in Fig. 1. While today's search engines support fact finding and link traversal well, they fail to equip scholars with a bird's eye view of the global structure and dynamics of scholarly knowledge and expertise. However, a more global view is needed to understand the structure and dynamics of science as it is conducted today, to detect emergent research frontiers (e.g., based on highly cited research articles), and for many other related tasks, e.g., setting research priorities.

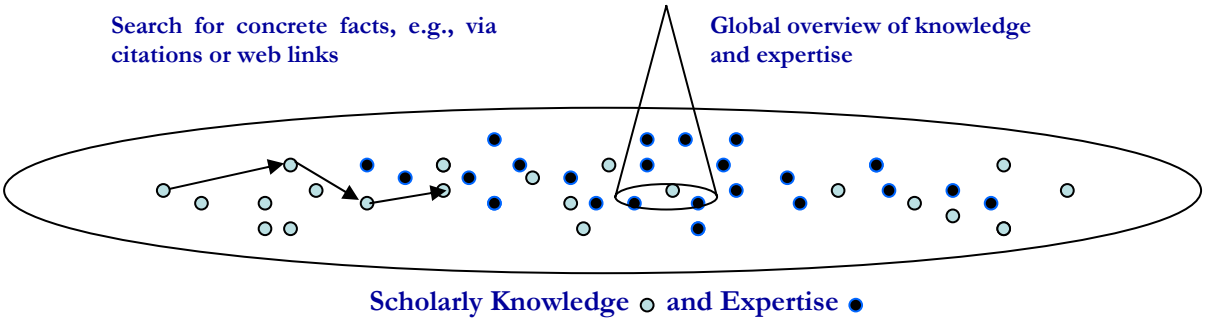

Figure 2: Search for concrete facts vs. gaining a global overview of the structure and evolution of science

[2] http://xanadu.com/

[3] This estimate was made based on our work on validating knowledge domain visualizations. Depending on the coverage of the knowledge domain map and the expertise of our subjects, 5-10 zooms of a 1:3 to 1:10 zooming factor are necessary until subjects recognize the set of papers that they are familiar with.

Visual interfaces to digital libraries (Börner & Chen, 2002) aim to help users mentally organize, electronically interact with, and manage large, complex information spaces. They apply powerful data analysis and information visualization techniques (Card et al., 1999) to transform data and information that are not inherently spatial into a visual form. The visualizations shift the user's mental load from slow reading to faster perceptual processes such as visual pattern recognition. Frequently, visual interfaces exploit human beings' powerful spatial cognition (Lakoff, 1987) and the method of loci – a mnemonic technique that originated with the ancient Greeks – to associate and attach any digital information, tool, or service to a spatial location or, using an identification tag, to other people. They have been successfully implemented and deployed for the PubMed database [4] by Antarcti.ca System Inc., and also in form of self organizing mapsfor astronomy and astrophysics articles[5].

Another major shortcoming of today's digital information spaces is the scarcity of social navigation cues (e.g., who is online, what resources are accessed frequently, etc.), making it difficult to find relevant resources and expertise or to collaborate. Research on social visualizations (Freeman, 2000; Donath et al., 1999; Erickson et al., 1999; Börner & Penumarthy, 2003) aims to show data about a person, illuminate relationships among people - even people they do not know - or visualize group activity to facilitate information access, collaboration, and decision-making.

People's 'information neighborhood' is not only made up of documents, but also - and perhaps more importantly - by people, including family, friends, neighbors, co-workers, and a shifting network of acquaintances (Haythornthwaite & Wellman, 1998). Having access not only to knowledge, but also to expertise – experts with deep knowledge about a subject matter – seems to be important for human 'information foraging' (Sandstrom, 1999; Pirolli & Card, 1999). Research on community knowledge portals aims to help participants capture, access, and manage knowledge and expertise created during their work process, to link community members to each other and to relevant content, and to offer personalized services tailored to the individuals and communities based on collaborative filtering (Mack et al., 2001).

A system that aims to support scholarly knowledge and expertise management should have an interface that is as easy to use as the Google search interface[6]. It should let users query not only for author names and paper titles, but also for the datasets used and the services (e.g., techniques, approaches, and/or software packages) applied to arrive at the results reported in the papers. In order to enable scholars to track the diffusion of scholarly knowledge via co-authorships and the citation of papers, the tool should also support users in the traversal of co-authorship and citation linkages. Visual interfaces might be employed to give users a global view of a dataset, a co-author network, or a paper-citation network. Usage data should be collected to support social navigation and to estimate the usefulness of data records and author contributions.

11.3 Semantic Association Networks

In *Needed – A National Science Intelligence and Documentation Center,* Eugene Garfield (1955b) wrote: "We may have ivory towers, but they are all connected by cables under the ground or by invisible channels in the air. Scientific progress – the accumulated effect of your individual contributions – depends upon a free flow of information, of thousands of minute facts, of millions of seemingly unrelated observations made and reported by scientists in diverse specialties."

[4] http://pubmed.antarcti.ca/start
[5] http://simbad.u-strasbg.fr/A+A/map.pl
[6] http://www.google.com/

This section describes *Semantic Association Networks* (SANs), a new means to interlink, access, and manage scholarly knowledge and expertise. It explains association based storage and access using SANs, their potential application for judging data quality and author reputation, and how they may be used to gain a more global view of knowledge and expertise. I start with a review of publication networks and publication-author networks and describe their extension to dataset-service-publication-author networks.

Publication Networks

Today, an increasing number of publishers and some search engines support the traversal of scholarly publications via citation linkages. The citation linkages are either supplied by hand or extracted automatically. CiteSeer (Giles et al., 1998) and the Google scholarly search engine[7] successfully support users in the traversal of automatically extracted citation networks.

Publication-Author Networks

Each publication also reports a set of authors that produced the publication. Similar to the automatic extraction of citation interlinkages, authors can be identified and interconnected via bi-directional 'co-authors_with' links, see Fig. 3. This makes possible the traversal of co-author networks and their associated publication results. Queries such as: "Retrieve all authors that collaborated with author x." can be executed.

Publication dataset can also be mined to determine the productivity, reputation, and topical expertise of an author. The number of papers published per time unit can be used as a measure for publication productivity. The quality of publication venues and the number of received citations are an indicator for the author's reputation. Keywords associated with an author's publications or the words appearing in the title, abstract, or full text of these publications can be used to identify topic coverage, topic changes, etc. Queries such as: "Retrieve all authors that have many publications in year x." or "Retrieve all authors that have a high reputation in area x." can be executed.

Information on co-author relationships can also be used to determine successful co-author teams (Börner et al., in press) and to answer queries such as: "Retrieve the set of authors that is most successful in area y."

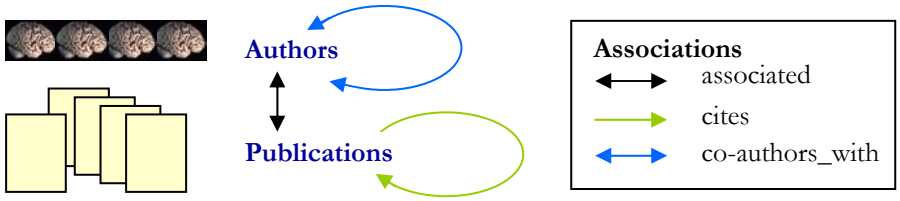

Figure 3: Publication-citation and associated co-author networks: 'Publications' are interconnected by 'cites' links. 'Authors' are connected by bi-directional 'co-authors_with' links. Each 'Publication' has an associated set of 'Authors'.

Data-Service-Publication-Author Networks

Today, diverse scientific communities create and contribute not only to repositories of scholarly publications, but also to repositories of datasets and services. Although a time and resource demanding enterprise, the creation of central databases gives researchers easy access to existing datasets.

For example, the National Library of Medicine provides access to the Online Mendelian Inheritance in Man (OMIM) database, a catalog of human genes and genetic disorders authored and edited by Dr. Victor A. McKusick and his colleagues at Johns Hopkins and elsewhere. OMIM interlinks gene data and

[7] http://scholar.google.com/

publication data based on manually identified correlations. Service repositories facilitate sharing, evaluation, and comparison of algorithms and software, and reduce the time and effort spent on repeatedly re-implementing algorithms. For example, the CERN Library (http://wwwinfo.cern.ch/asd/) provides a large collection of general purpose programs maintained and offered in both source and object code. Researchers and students are encouraged to use this code repository rather than their own code. Aside from saving users time and effort, the repository code is more likely to be correct after having been tested by many other people.

Plug-in based software frameworks (Penumarthy et al., Submitted) that integrate algorithms written in diverse programming languages, that support a multitude of file formats and offer menu driven interfaces, appear to serve interdisciplinary user communities well.

If data and services are associated with the publications and authors that report them, then new means of search and association discovery become possible (see Fig. 4). Queries such as: "Retrieve all authors that used service x.", "Retrieve all papers that report results from dataset y.", "Retrieve all services that were used with dataset y." can be executed.

The availability of data, service and publication repositories also eases the replication of results reported in papers, the application of existing services to new data or the analysis of existing data with new services. Service repositories support the examination of algorithms at the code level – as opposed to the pseudo-code level reported in papers.

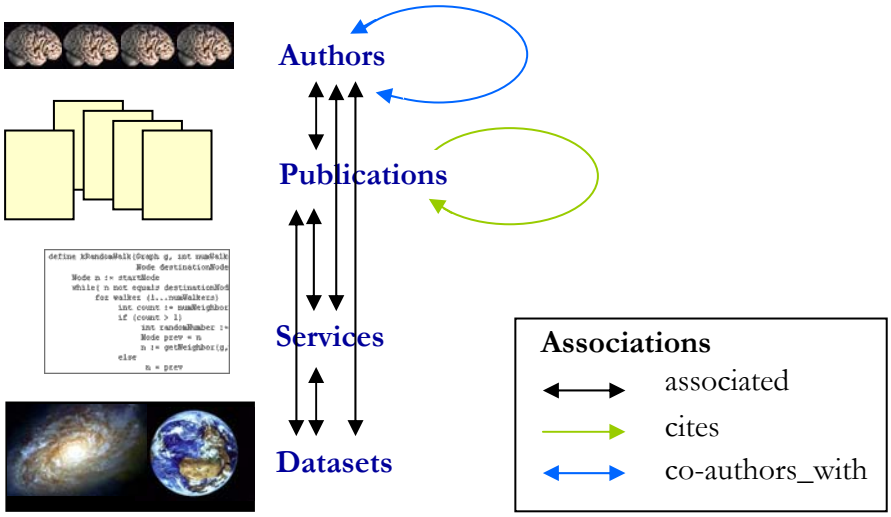

Figure 4: 'Publications' associated with 'Authors', 'Services' and 'Datasets'.

Users of association networks would contribute publications that also report datasets and services used. If they used new datasets and/or services they would be required to submit those as well (see Fig. 4). Note that Journals such as *Science* and *Nature* already require authors to provide access to datasets and services before a paper is published. The *National Science Foundation* and the *National Institutes of Health* also promote and financially support the development of datasets and service repositories.

From the submitted publications, information on 'Authors' and the 'associated' links to 'Authors', 'Services' and 'Datasets' can be derived automatically, as can 'cites' and 'co-authors_with' relationships.

A major effort will be the identification of identical authors, publications, services and datasets. Authors may report their names in different ways, e.g., with or without a middle name. They may change their name due to marriage. Publishers might translate and spell foreign names in different ways. Publications might report nearly or completely identical results. Services, e.g., clustering algorithms, are developed by diverse communities and most likely are given different names. It is only at the code level and/or input/output level that they can be compared. Identical datasets might exist in different formats and with different names and metainformation.

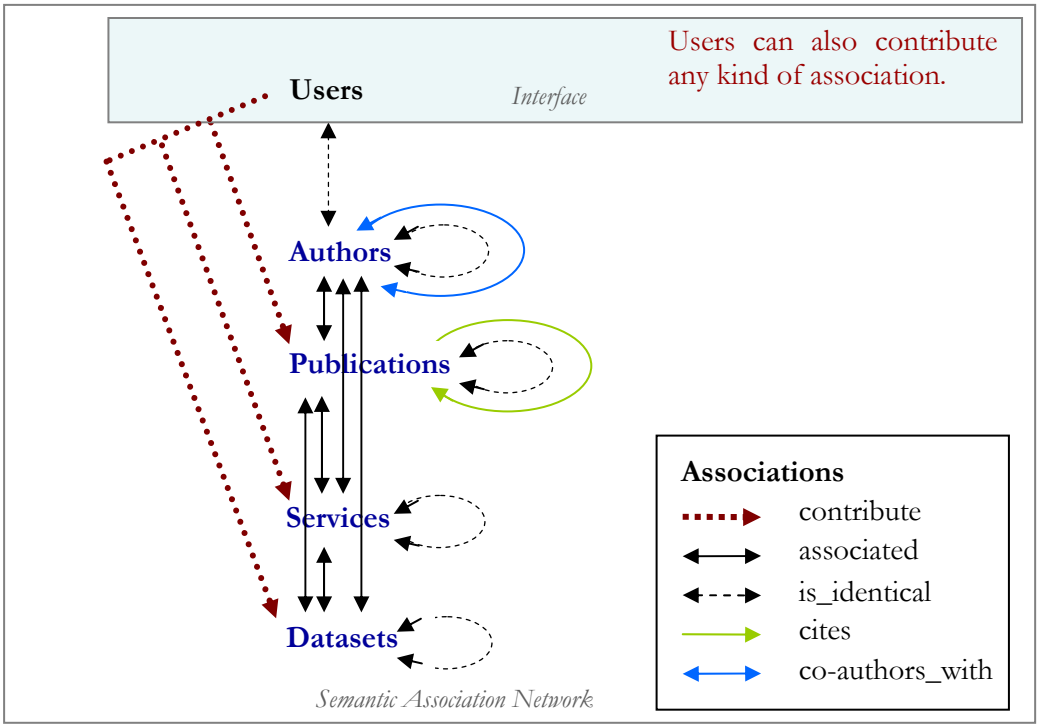

Figure 5: *Semantic Association Networks* interconnect 'Datasets', 'Services', 'Publications', 'Authors' and 'Users' via diverse associations.

'Is_identical' associations are employed to denote if two authors are the same, if two or more publications report the same result, if a set of services has identical input/output behavior, or if a set of data is identical (see Fig. 5). Identity can be established using automatic means or could be supplied by users of the system.

User activity would be logged and the number of accesses would be used to estimate the value and quality of any of the data records. Resources that are accessed or cited more often or are published in higher reputation journals might be given higher quality placement. Reputation would be given to authors/donators of high quality resources.

Visual interfaces should be implemented to provide users of SANs with a global, bird's eye view of the data that is available at their finger tips. Visualizations would also reveal and help users understand the complex associations among the different data entries. They would help users identify areas of high activity and/or growth, understand what data is accessed most often, understand what data is widely regarded as most correct, etc.

Some very promising examples of effective interfaces to publication data are already available. *VisualLink*, developed by Xia Lin, Howard D. White, and Jan Buzydlowski, provides a global view of author association networks by analyzing and visualizing the number of times authors are cited together (Lin et al., 2003). The *HistCite* software developed by Eugene Garfield and colleagues generates global maps of paper-citation networks in support of a local and global examination of highly cited papers and their interlinkages (Garfield, 2004).

Semantic Association Networks in RDF Representation

Semantic Web technology such as the W3C's Resource Description Framework (RDF) (http://www.w3.org/RDF/) can be applied to define and represent *Semantic Association Networks* in a way such that automatic association discovery, retrieval, and management become possible.

RDF is a very simple data model for representing information about online resources. Each resource (any identifiable thing, including things that may not be directly retrievable on the Web) has a *Uniform Resource Identifier*, or *URI.* An URI can have simple properties and property values. To ease adoption, a relatively small subset of RDF will be applied to encode the semantics of SANs. Fig. 6 shows an exemplary RDF presentation of SANs that is explained subsequently.

The RDF representations of SAN nodes are discussed first. 'Author', 'Paper', 'Service' and 'Dataset' nodes in the SAN are defined as subclasses of 'Entity' with properties 'entity_created_on' and 'times_accessed'. Hence all their instantiations have an attribute value for 'entity_created_on' that indicates their time of creation as well as a value for 'times_accessed' denoting how often they have been accessed by a user. Each instantiation of any of the subclasses has a unique ID. Multiple implementations of one service might have different IDs – until their identity is discovered. A class 'Association' is defined with properties 'association_created_on' and 'times_occurences' to specify the time at which SAN links are generated and how often they occur (e.g., how often a pair of authors co-authored). Class 'User' has a property 'contributed' that provides information on what 'User" contributed what 'Entity' or 'Association'.

SAN links such as 'associated', 'co-author_of', 'cites', 'is_identical_author', 'is_identical_paper', 'is_identical_service', 'is_identical_dataset' are represented as subclasses of the class 'Association'. The bi-directional 'associated' links exist among subclasses of type 'Entity': 'Datasets', 'Services', 'Publications', and 'Authors'. Only 'Papers' can be linked via directed 'cites' links and only 'Authors' can be linked via bi-directional 'co-authors_with' links. Bi-directional 'is_identical_*' links exist among subclass instantiations of the same type as well as among 'Users' and 'Authors'. Directed 'contributes' relations exist between users and any subclass of 'Entity' or 'Association'.

Note that the proposed RDF metadata would be assigned automatically, not manually. Users would login to a scholarly repository using a unique author identifier. If they access an 'Entity' then the 'times_accessed' property of this entity would be updated automatically. If they submit an 'Entity' or 'Association' then a '*_created_on' property would be generated and their user ID would be connected via a 'contributes' property to the new entity and/or association. The user never sees any RDF tags, but their availability makes new means of search, interlinkage, and management possible.

Note that the RDF specification given in Fig. 6 can be easily extended to incorporate other data such as 'Comments' or an 'evaluation score' 'contributed' by an 'User' for any 'Entity'. The former could be implemented as another subclass of 'Entity'. The latter could be encoded as an additional property of 'Entity'. It could also be beneficial to have subclasses of class 'Publication' denoting that a certain publication is a 'Paper', 'Patent', or 'Grant' or to tag a paper as being a 'Journal Paper', 'Book Chapter', etc.

```
<!—Define class Entity and all its properties -->
```

```
<rdfs:Class rdf:ID="Entity"/>
<rdf:Property rdf:ID="entity_created_on">
    <rdfs:domain rdf:resource="#Entity"/>
</rdf:Property>
<rdf:Property rdf:ID="times_accessed">
    <rdfs:domain rdf:resource="#Entity"/>
</rdf:Property>

<!—Define subclasses of Entity-->
<rdfs:Class rdf:ID="Author">
    <rdfs:subClassOf rdf:resource="#Entity"/>
</rdfs:Class>
<rdfs:Class rdf:ID="Paper">
    <rdfs:subClassOf rdf:resource="#Entity"/>
</rdfs:Class>
<rdfs:Class rdf:ID="Service">
    <rdfs:subClassOf rdf:resource="#Entity"/>
</rdfs:Class>
<rdfs:Class rdf:ID="Dataset">
    <rdfs:subClassOf rdf:resource="#Entity"/>
</rdfs:Class>

<!—Define class Association and all its properties -->
<rdfs:Class rdf:ID="Association"/>
<rdf:Property rdf:ID="association_created_on">
    <rdfs:domain rdf:resource="#Association"/>
</rdf:Property>
<rdf:Property rdf:ID="times_occurences">
    <rdfs:domain rdf:resource="#Association"/>
</rdf:Property>

<!—Define subclasses of Association-->
<rdf:Property rdf:ID="associated">
    <rdfs:subClassOf rdf:resource="#Association"/>
    <rdfs:domain rdf:resource="#Entity"/>
    <rdfs:range rdf:resource="#Entity"/>
</rdf:Property>
<rdf:Property rdf:ID="co-authors_with">
    <rdfs:subClassOf rdf:resource="#Association"/>
    <rdfs:domain rdf:resource="#Author"/>
    <rdfs:range rdf:resource="#Author"/>
</rdf:Property>
<rdf:Property rdf:ID="cites">
    <rdfs:subClassOf rdf:resource="#Association"/>
    <rdfs:domain rdf:resource="#Paper"/>
    <rdfs:range rdf:resource="#Paper"/>
</rdf:Property>
<rdf:Property rdf:ID="is_identical_author">
    <rdfs:subClassOf rdf:resource="#Association"/>
    <rdfs:domain rdf:resource="#Author"/>
    <rdfs:range rdf:resource="#Author"/>
    <rdfs:range rdf:resource="#User"/>
</rdf:Property>

<!—define rdf:Property "is_identical_paper", "is_identical_service",
 "is_identical_dataset" analogously to rdf:Property "is_identical_author"-->

<!—Define class User and its property -->
<rdfs:Class rdf:ID="User"/>
<rdf:Property rdf:ID="contributes">
    <rdfs:domain rdf:resource="#User"/>
    <rdfs:range rdf:resource="#Association"/>
    <rdfs:range rdf:resource="#Entity"/>
</rdf:Property>
</rdf:RDF>
```

Figure 6: *Semantic Association Networks* in RDF representation

To implement the proposed *Semantic Association Networks*, scientific communities, institutions and companies will need to record, make available, and preserve not only scholarly papers, but also datasets and software services. In addition, authors need to be encouraged to report used datasets and applied services in their publications. This will provide the basis for the semi-automatic generation of SANs.

Most likely, a combination of automatic citation indexing (Giles et al., 1998) and association discovery, and simple web forms via which registered users can correct or add links manually will be most successful in acquiring accurate and comprehensive SANs.

Initially, SANs and user registration could be hosted and managed centrally. Over time, SAN registries would need to become decentralized to improve access time and error tolerance. The growth and usage of the semantic association networks should be monitored extensively. Resulting usage data could be employed to optimize the SANs, e.g., to mirror highly accessed resources to improve access time.

11.4 Implementing SANs: Opportunities and Challenges

The implementation of SANs provides a number of potentially very powerful new means to access and correlate scholarly knowledge and expertise. SANs could help to create a scholarly 'interoperability substrate' to design value-adding service that are not 'vertically' locked. For example, they provide the data needed for the generation of effective visual interfaces to digital libraries or the design of global maps of science that report the structure and evolution of science.

Note that SANs are envisioned to serve all areas of science, interconnecting and hence helping to cross-fertilize interdisciplinary research. Given that everybody with Internet access – including school kids or hobby researchers – would be able to use SANs to gain access to first rate knowledge and expertise, SANs can be seen as a means of mass education and empowerment.

Widespread availability of SANs would also constitute a great testbed application for diverse scientific communities that aim to develop novel ways to store, preserve, integrate, correlate, access, analyze, map or interact with data.

Subsequently, I review the diverse challenges that one would face when trying to implement SANs on a global scale.

Social Challenges

As mentioned above, appropriate incentive structures need to be implemented to make authors donate not only papers, but also datasets and services used.

Standards will have to be developed to uniquely, persistently, and globally identify the content of a data record (e.g., papers, authors, datasets and services). The *Digital Object Identifier*[8] (DOI) system might be used and extended to also cover authors, datasets and services. Another rich source of information on bibliographic, holdings, authority, classification, and community information is the MARC[9] (MAchine-Readable Cataloging) data format that emerged from a Library of Congress[10] led initiative. MARC records (which became USMARC in the 1980s and MARC 21 in the late 1990s) have been extensively used by most libraries for the last 30 years. The MARC author name authority records might provide a useful nucleus for the creation of a unique author identifier. While these author authorization files identify authors via a fixed name, date of birth, etc., I believe that an author identifier should not contain the name

[8] http://www.doi.org/
[9] http://www.loc.gov/marc/
[10] http://www.loc.gov/

of the author (as it might change due to marriage, etc.), the date of birth, social security numbers, or passport numbers.

Other social challenges relate to data protection, privacy concerns, legitimacy via content contribution and evaluation by distributed subject and professional teams, and sustainable resource models. The scale-free topology of scientific networks poses serious challenges with regard to multilingualism, preservation of diverse traditions, views, and approaches as only a minority of sources and experts is highly visible and accessible while the vast majority is too weakly connected to be seen.

Technological Challenges

If scholars can be persuaded to provide access to datasets and services and to report the usage of datasets and services in the papers they publish, then automatic means can be employed to associate data/services, papers, and their authors by means of semantic association networks. Major technical challenges comprise the federation of data from different databases, dealing with all kinds of data quality issues (e.g., multiple formats, misspellings, omissions, etc.), planning for sustainability, robustness, support of heterogeneous hardware, and the preservation of datasets and services as technology and data formats evolve.

I do not foresee a shortage in terms of computing power, data storage or bandwidth. Compute power is doubling almost every year (Djorgovski, 2004) and theoretically, this trend can continue for 600 more years (Krauss & Starkman). The performance/price of disk-based data storage is improving even faster than the performance/price of computing power – at a factor of about 100 over the last ten years (Hayes, 2002). Bandwidth is increasing at a much slower pace. Global efforts such as the TeraGrid[11] - a multi-year effort to build and deploy the world's largest, most comprehensive, distributed infrastructure for open scientific research - will soon interconnect major databases and services at high bandwidth. However, the cost effective interconnection of customer sites with comparatively low amounts of traffic to major bandwidth hubs, also called the 'last mile problem', is unresolved.

Given the compute power, data storage capacity, and bandwidth available today it is surprising to see how old-fashioned our current means to access and manage scholarly knowledge and expertise are.

Funding Acquisition Challenges

The development of standards and the implementation of SANs will require funding. The effort to implement SANs might be best compared with the implementation of the diverse Citation Indexes provided by the Institute of Scientific Information (ISI) or the implementation of Google. However, the dollar amounts spent on those two projects are not available to us.

Instead, I report the effort spent on Douglas Lenat's Cyc[12] project – most likely the world's largest common sense knowledge base in existence today. The Cyc project was funded over 20 years with $25 million Artificial Intelligence research dollars. It is a 600 person per year effort that assembled a knowledge base containing 3 million rules of thumb that the average person knows about the world, plus about 300,000 terms or concepts nodes (a typical person is assumed to have about 100 million concepts).

The proposed SANs differ from Cyc in that they do not require the careful manual compilation and encoding of logical rules which interconnect concept nodes. Instead, a semi-automatic process is proposed to cross-correlate existing data, services, publication, and author databases. The interlinkage of all scholarly knowledge might very well exceed the resources spent to create Cyc but it would benefit all of science.

[11] http://www.teragrid.org/
[12] http://www.cyc.com/

Where to Start?

Physics and astronomy are two scholarly communities that keep particularly good track of their data, services, and publications. International efforts such as the International Virtual Observatory[13] aim to develop standards for data and interfaces as well as for software packages, source code libraries, and development tools (Williams et al.). The American Physics Society has *Physical Review Letters* publication data available in full text from 1893 until today (OCR'ed full text exists for everything from about 1995 back). Physicists and astronomers use the arxiv.org e-print archive extensively for timely scholarly publication. The physics and astronomy domains would be good candidates for implementing prototype SANs.

11.5 Concluding Remarks

Eugene Garfield, when proposing a *Unified Index to Science* wrote: "Grandiose schemes always meet with excessive resistance, not because they are impossible to achieve, but because there are only a few with sufficient persistence to materialize their dreams and even fewer to carry them out. Ultimately, most large endeavors must fall by the wayside, to be replaced by others. However, their value at a particular stage of history cannot be disputed." (1959), p. 468.

This chapter described *Semantic Association Networks*, a new means to tag, interconnect, and manage scholarly datasets, services, results, and expertise as a means to keep better records of our collective scientific knowledge, expertise, and progress.

I conclude this chapter with a vision for the future of scholarly publishing that might supersede today's writing, peer-review, and publication of papers. This process becomes unmanageable as the flood of information is increasing and there are no automatic means to extract and compile knowledge summaries for paper collections.

Somewhere in the not too distant future, reporting a scholarly result might not involve writing a paper. Instead, scholars may add a "knowledge nugget node" or an "association link" to a complex semantic association network of humanity's knowledge – some Bloggers are already practicing this today. The nodes in this network will describe tangible objects (e.g., a pottery piece found at a certain place by an archeologist together with information about its origin and intermediate positions/usages up to today) or intangible objects (e.g., a formula). Links will represent associations of diverse types (e.g., causal ones such as "ozone is created electrically in nature during active thunderstorms"). Each node and each link would have information on who added, modified or deleted it. A scholar's reputation would depend on the number of nodes and/or links s/he contributed and their usefulness for humanity.

Scholars would navigate, mine, and add to this vast network of humanity's knowledge using the information associated with nodes and links and data about the usage of nodes and links. If a scholar adds a new "knowledge nugget" or "association link" the network would be 're-compiled' with the new node leading to

- the identification of redundant data/algorithm/results,
- a confirmation of existing facts improving the correctness of humanity's knowledge,
- a confirmation of the novelty of a fact increasing humanity's knowledge, or
- a conflict with existing facts.

In the latter case, either the new nugget or link is wrong, or other nodes/links are wrong, alternative interpretations/views are acceptable, or the entire conceptualization is not working and a scientific revolution in Kuhn's sense (1962) is needed.

[13] http://us-vo.org/

Acknowledgements

I would like to thank Kevin Boyack, Joe Futrelle, Stacy Kowalczyk, Deborah MacPherson, Ketan K. Mane, Shashikant Penumarthy, Bonnie DeVarco, and Elijah Wright for insightful comments on a draft of this paper. Shashikant Penumarthy inspired and collaborated on the representation of SANs in RDF format. Eugene Garfield, Peter A. Hook, and Mark Meiss provided helpful pointers. This work was supported by a National Science Foundation CAREER grant under IIS-0238261.

References

Borgman, C. 2000. Scholarly communication and bibliometrics Revisited. In: *The Web of Knowledge* (Ed. by Cronin, B. & Atkins, H. B.), pp. 143–162. Medford, NJ: Information Today.

Börner, K. & Chen, C. 2002. Visual Interfaces to Digital Libraries. In: *LNCS*: Springer Verlag.

Börner, K., Chen, C. & Boyack, K. 2003. Visualizing Knowledge Domains. In: *Annual Review of Information Science & Technology* (Ed. by Cronin, B.), pp. 179-255. Medford, NJ: Information Today, Inc./American Society for Information Science and Technology.

Börner, K., Dall'Asta, L., Ke, W. & Vespignani, A. in press. Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams. *Complexity*.

Börner, K. & Penumarthy, S. 2003. Social Diffusion Patterns in Three-Dimensional Virtual Worlds. *Information Visualization*, **2,** 182-198.

Bush, V. 1945. As we may think. *The Atlantic Monthly*, **176,** 101-108.

Card, S., Mackinlay, J. & Shneiderman, B. 1999. Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann.

Djorgovski, S. G. 2004. Virtual Observatory, Cyber Science, and the Rebirth of Libraries. *Available at: http://www.arl.org/forum04/djorgovski.html, Accessed 11/16/2004*.

Donath, J. S., Karahalios, K. & Viegas, F. 1999. Visualizing conversation. *Journal of Computer Mediated Communication*, **4**.

Engelbart, D. C. 1963. A Conceptual Framework for the Augmentation of Man's Intellect. In: *Vistas in Information Handling* (Ed. by Howerton, P. W. & Weeks, D. C.), pp. 1-29. Washington D.C.: Spartan Books.

Erickson, T., Smith, D. N., Kellogg, W. A., Laff, M., Richards, J. T. & Bradner, E. 1999. Socially Translucent Systems: Social Proxies, Persistent Conversation, and the Design of "Babble". In: *Proceeding of the CHI 99 Conference on Human Factors in Computing Systems: The CHI is the limit*, pp. 72 - 79: ACM Press.

Freeman, L. C. 2000. Visualizing Social Networks. *Journal of Social Structure*, **1**.

Garfield, E. 1955a. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, **122,** 108-111.

Garfield, E. 1955b. Needed - A National Science Intelligence and Documentation Center. Presented at the "Symposium on Storage and Retrieval of Scientific Information" of the annual meeting of the American Association for the Advancement of Science in Atlanta, Georgia. Available at http://www.garfield.library.upenn.edu/papers/natlsciinteldoccenter.html, Accessed 10/18/2004.

Garfield, E. 1959. A Unified Index to Science. In: *International Conference on Scienctific Information*, pp. 461-474. Washington, D.C.: National Academy of Sciences, Available at http://books.nap.edu/html/sci_inf_1959/461-474.pdf, Accessed 10/18/2004.

Garfield, E. 2004. Historiographic mapping of knowledge domains literature. *Journal of Information Science*, **30,** 119-145.

Giles, C. L., Bollacker, K. & Lawrence, S. 1998. CiteSeer: An automatic citation indexing system. In: *Digital Libraries 98 - The Third ACM Conference on Digital Libraries* (Ed. by Ian Witten, R. A., and Frank M. Shipman III), pp. 89–98. Pittsburgh, PA: ACM Press.

Hayes, B. 2002. Terabyte Territory. *American Scientist*, **90,** 212–216.

Haythornthwaite, C. & Wellman, B. 1998. Work, friendship, and media use for information exchange in a networked organization. *Journal of the American Society for Information Science*, **49,** 1101-1114.

Henry, G. 2003. On-line publishing in the 21-st Century: Challenges and Opportunities. *D-Lib Magazine*, **9**.

Krauss, L. M. & Starkman, G. D. Universal Limits on Computation. *arxiv: astro-ph/0404510*.

Kuhn, T. S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lakoff, G. 1987. *Women, Fire, and Dangerous Things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Lin, X., White, H. D. & Buzydlowski, J. 2003. Real-time author co-cocitation mapping for online searching. *Information Processing and Management*, **39,** 689-706.

Lynch, C. 2003. Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *ARL Bimonthly Report 226***,** Available at http://www.arl.org/newsltr/226/ir.html, Accessed 10/19/2004.

Mack, R., Ravin, Y. & Byrd, R. J. 2001. Knowledge portals and the emerging digital knowledge workplace. *Ibm Systems Journal*, **40,** 925-955.

Penumarthy, S., Börner, K. & Herr, B. Submitted. Information Visualization Cyberinfrastructure Software Framework. *Information Visualization*.

Pirolli, P. & Card, S. 1999. Information foraging. *Psychological Review*, **106,** 643-675.

Price, D. J. D. 1965. Networks of scientific papers. *Science*, **149,** 510-515.

Sandstrom, P. E. 1999. Scholars as subsistence foragers. *Bulletin of the American Society for Information Science*, **25,** 17-20, Available at http://www.asis.org/Bulletin/Feb-99/sandstrom.html, Accessed 10/18/2004.

Scholtz, J. 2000. DARPA/ITO Information Management Program Background.

Shiffrin, R. 2004. Scientists seek 'map of science'. BBC News, Available at http://news.bbc.co.uk/1/hi/sci/tech/3608385.stm, Accessed 10/24/2004.

Shiffrin, R. & Börner, K. 2004. *Mapping Knowledge Domains, Proceedings of the National Academy of Sciences*. (Suppl. 1), Volume 101.

Smith, M., Bass, M., McClellan, G., Tansley, R., Barton, M., Branschofsky, M., Stuve, D. & Walker, J. 2003. DSpace: An Open Source Dynamic Digital Repository. *D-Lib Magazine*, **9**.

Van de Sompel, H., Payette, S., Erickson, J., Lagoze, C. & Warner, S. 2004. Rethinking Scholarly Communication: Building the System that Scholars Deserve. *D-Lib Magazine*, **10**.

Wells, H. G. 1938. *World Brain*. Garden City, NY: Doubleday, Doran.

Williams, R., Moore, R. & Hanisch, R. A Virtual Observatory Vision based on Publishing and Virtual Data, Available at http://us-vo.org/pubs/files/VO-vision.pdf, Accessed on 11/10/2004.

Williams, R., Moore, R. & Hanisch, R. 2003. A Virtual Observatory Vision based on Publishing and Virtual Data. Available at http://bill.cacr.caltech.edu/usvo-pubs/files/VO-vision.pdf, Accessed 10/19/2004.