Using hybrid search and query for e-discovery identification

Dave Grosvenor, Andy Seaborne

HP Laboratories

HPL-2009-155

Keyword(s):
SPARQL, e-discovery, identification, hybrid search and query

Abstract:

We investigated the use of a hybrid search and query for locating enterprise data relevant to a requesting party's legal case (e-discovery identification). We extended the query capabilities of SPARQL with search capabilities to provide integrated access to structured, semi-structured and unstructured data sources. Every data source in the enterprise is potentially with in the scope of E-discovery Identification. So we use some common enterprise structured data sources that provide product and organizational information to guide the search and restrict it to a manageable scale. We use hybrid search and query to conduct a rich high-level search, which identifies the key people and products to coarsely locate relevant data-sources. Furthermore the product and organizational data sources are also used to increase recall which is a key requirement for e-discovery Identification.

External Posting Date: July 6, 2009 [Fulltext] Approved for External Publication
Internal Posting Date: July 6, 2009 [Fulltext]

Using hybrid search and query for e-discovery identification

Dave Grosvenor[1], Andy Seaborne[1]

[1]Hewlett-Packard Laboratories, Bristol
{dave.grosvenor, andy.seaborne}@hp.com

Abstract. We investigated the use of a hybrid search and query for locating enterprise data relevant to a requesting party's legal case (e-discovery identification). We extended the query capabilities of SPARQL with search capabilities to provide integrated access to structured, semi-structured and unstructured data sources. Every data source in the enterprise is potentially with in the scope of E-discovery Identification. So we use some common enterprise structured data sources that provide product and organizational information to guide the search and restrict it to a manageable scale. We use hybrid search and query to conduct a rich high-level search, which identifies the key people and products to coarsely locate relevant data-sources. Furthermore the product and organizational data sources are also used to increase recall which is a key requirement for e-discovery Identification.

Keywords: SPARQL, e-discovery, identification, hybrid search and query

1 Introduction

E-discovery is the process of collecting, preparing, reviewing, and producing electronic documents in a variety of criminal and civil actions and proceedings [1]. In this paper we address the problems of scale and recall in the *identification* stage of e-discovery which is responsible for learning a coarse location of data relevant to a legal case. There are two components to our approach to these problems. The first component was to add search directives to SPARQL [2] to give two different information retrieval models in a hybrid search and query. This gives integrated access to both structured and unstructured data sources in the enterprise. However the second component was to exploit some common product and organizational data sources to both guide the searches to cope with scale, and to increase recall.

This paper is organized as follows:

- First we give an extended introduction with: an overview of e-discovery; an explanation of identification and where it fits within this process; we then introduce the problems posed by identification; before ending with a discussion of related work.

- Second we discuss our approach to the problems posed by identification, discussing: our addition of search directives to SPARQL; and our use of product and organizational information.
- Third we examine a hypothetical patent violation e-discovery case and give examples on the use of hybrid search and query.
- Finally we give our conclusions on the investigation.

E-discovery

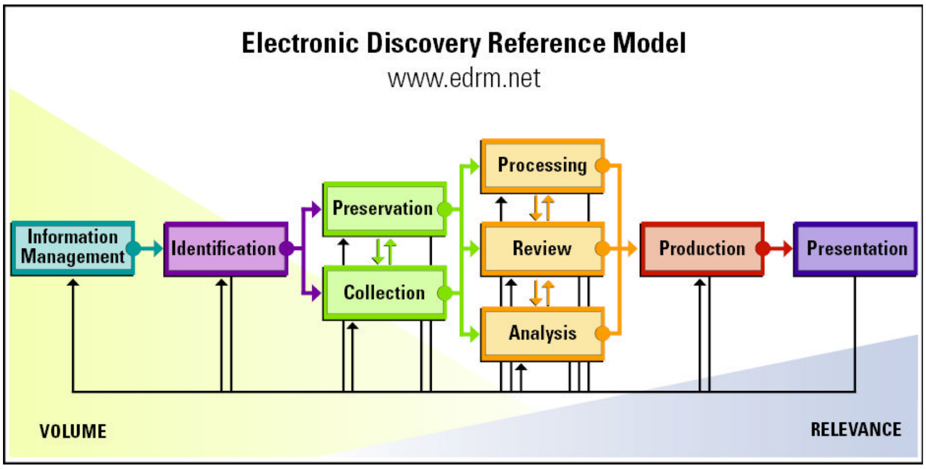
E-discovery is a new issue for enterprises created in 2006 by the Federal Rules for Civil Procedure (FRCP) [3] in the US which legally require enterprises:

- To disclose the identities of all individuals likely discoverable information relevant to a legal case.
- To either provide a copy, or the location and description of all Electronically Stored Information (ESI) relevant to a legal case.

The courts can potentially impose punitive damages on an enterprise for a failure to comply. The ESI includes both structured and unstructured information within the enterprise which specifically includes e-mails, web-pages, word processing files, and databases stored in the memory of computers, magnetic disks (such as computer hard drives and floppy disks), optical disks (such as DVDs and CDs), and flash memory (such as "thumb" or "flash" drives). Currently email is the most important source of discoverable information (80-90% according to a Magistrate Judge Survey [4]). For example email analysis was used extensively in the ENRON case and is a subject of the TREC legal track [5].

E-discovery requests can be initiated by arbitrary legal cases which makes it difficult to prepare for in advance. According to the "Magistrate Judge Survey" [4] only a few types of case are responsible for most of the E-discovery issues. These types of cases were: Individual plaintiffs in Employment cases; General Commercial cases, Patent or Copyright cases, Class Employment actions, and Product Liability cases. The Federal Judicial Centre provides examples of e-discovery requests [8], preservation orders [9], and 'meet and confer' forms [10]. The Sedona conference [11] has been influential in the development of approaches to e-discovery. In this paper we examine a fictitious patent violation case.

E-discovery is a complex guided search task conducted over a long duration (perhaps several months) by an expert team consisting of legal and IT experts acting on behalf of the disclosing party. The *Electronic Discovery Reference Model* (EDRM) [6] has been developed to explain and position different E-discovery products.



The reference model decomposes the e-discovery process into:

- An I*nformation Management* stage responsible for all preparation prior to an E-discovery request, including records management and policies.
- An *Identification* stage responsible for learning the location of the data relevant to the discovery request.
- A *Preservation* stage responsible for ensuring the ESI identified in the previous step is protected from inappropriate alteration or destruction.
- A *Collection* step responsible acquiring the identified ESI.
- A *Processing* stage where the volume of data is further reduced and converted into a suitable format for *Review* and *Analysis*.
- A *Review* stage where a disclosing party's legal team sorts out both the responsive documents to produce, and the privileged documents to withhold.
- An *Analysis* stage where the collection of materials is evaluated to determine relevant summary information, such as key topics of the case, important people, specific vocabulary and jargon, and important individual documents.
- A *Production* stage where the ESI is delivered to the requesting party in both an appropriate form and by an appropriate delivery mechanism.
- A *Presentation* stage where the ESI produced is displayed before legal audiences (dispositions, hearings, trials, etc..).

Search technology is used for many stages of the reference model [7]. However e-discovery is not just an enterprise search application because: firstly the FRCP requires the identity of every person likely to hold relevant information, and so not only documents are returned; and secondly the emphasis in E-discovery is on returning all potentially responsive data whereas enterprise search returns a selection of the documents most likely to be relevant to a request. Precision and recall are two measures of information retrieval performance [12] that are commonly used.

- Precision is the fraction of retrieved documents which are relevant.

- Recall is the fraction of the relevant documents which have been retrieved.

Enterprise search is primarily concerned with precision whereas e-discovery is concerned with recall.

Identification

The *identification* stage of the e-discovery reference model poses several problems:

- The scale of data which is potentially retrievable during identification.
- The requirement for high recall.
- The characterization of relevance to an e-discovery request.

During identification any person, product, organization and data-source within the organization is potentially within its scope. This scope makes the brute force use of search technology more difficult during identification. An output of this stage is a manageable selection of the enterprise data that is potentially relevant to the E-discovery request. Only the data selected in identification is subjected to the more detailed review and analysis during of the later e-discovery stages.

The recall obtained during identification provides an upper bound for the overall recall during e-discovery and so high recall during identification is very important. As a result Identification casts a broad net in characterizing what constitutes relevant data. Identification characterizes relevant data by: identifying and interviewing the key witnesses and custodians of data sources; identifying the time frame relevant to the e-discovery request; identifying keyword lists, special language, jargon and technical terms relevant to the request; mapping out the disclosing company's data sources.

It usual for identification to have a strong manual element and it can be completely manual. It begins with interviews of potential custodians which are performed by the E-discovery team (custodian-led identification). Identification begins with key players because data of individuals who played a central role in the dispute are likely to contain the majority of information relevant to the dispute. Hybrid search and query supports rich model of the identification search task because it allows retrieval not only of documents but retrieval of people, organizations and products.

Related work

There are many approaches to hybrid search and query to which we provide some brief pointers. Research has aimed at search-like retrieval to access structured data, providing keyword search, imprecise queries, top-k selection and ranking [13]. Similarly other research has aimed at more query-like retrieval from unstructured data, using information extraction to provide: querying of unstructured text [14] and searches returning entities rather than just documents [15]. This research simply underlines the value of both models of information retrieval and our approach has been to make use of both. This pragmatic approach has been followed previously, such as in the WSQ/DSQ work [16] which combines the query facilities of traditional databases with existing search engines on the Web. WSQ/DSQ leverages both results from Web searches to enhance SQL queries over a relational database, and uses information stored in the database to enhance and explain Web searches. Parametric search (such as supported by Lucene [17] and most enterprise search engines) provides a similar capability to hybrid search and query by both associating metadata fields with each document that is indexed, and allowing queries to retrieve their value and to restrict searches to particular values.

There is closely related work [18][19][20][21] which adds full text search capability to SPARQL. They all are concerned search the literal strings in RDF datasets. The Sesame like operator [18] simply filters the results using regular expression matches on the literals in the result sets. Virtuoso [21] provides a system of rules for controlling which RDF triples are indexed. We use an ARQ mechanism for extending SPARQL that allows us to access arbitrary indexes and are not restricted to RDF datasets. There is additional support for both full text search and selectively indexing RDF with Lucene [17].

There are standard approaches to increasing recall using domain knowledge, such as query expansion [22] and spreading activation [23]. Both are automatic means of obtaining more responses to an original query using domain knowledge. This is important for e-discovery identification. Query expansion operates in the query space and transforms the query using the domain knowledge and co-occurring terms to find related or more general search terms and constraints. Spread activation operates in the result space and uses the initial results as seeds that are used to activate other related concepts during a propagation phase. We use neither technique, but we use the product catalog and the organizational data to identify related people and products which are used both in additional queries, and to generate further result.

2 Our approach to hybrid search and query

In this section we both motivate our use of SPARQL with search directives, and explain how search directives are evaluated in a hybrid query.

Motivation

Our approach is to exploit both structured and unstructured data sources for e-discovery which requires some form of information integration. The semantic web provides useful technology for integrating information across structured and unstructured data sources. The schema-less semi-structured RDF model provides a common data model suited for heterogeneous information integration. The RDFS/OWL ontology languages provide a way to describe the conceptual model behind information sources separate from the low level details of storage schemas. SPARQL gives us a query language suited to the data model with explicit support for cross-source query.

An RDF approach to integration provides a low cost of entry you can query and navigate RDF instance data without the need for semantic integration. This is important as e-discovery potentially requires ad hoc integration to bring together data sources for the particular legal request that would not be used together during the normal operation of the business.

This gives a high-level justification for the use of SPARQL to access structured data sources for E-discovery. Whilst RDF can be used to store the results of information extraction performed on unstructured documents the predominant means of retrieving unstructured documents uses a search retrieval model. Thus pragmatically the introduction of search directives into the SPARQL enables us to access search based data sources available in the enterprise.

SPARQL

SPARQL [2] is a standard query language, protocol and XML result set format defined by a W3C working group. It became a W3C recommendation in January 2008. A SPARQL query consists of a graph pattern and a query form (one of `SELECT`, `CONSTRUCT`, `DESCRIBE`, `ASK`). The simplest graph pattern is the basic graph pattern (BGP), a set of triple patterns that must all be matched. This is an extension point of the language and we utilize it to add semantic relationships which are not directly present in the data. In particular, we use other indexing technologies, such as free-text indexes to relate text query strings with document URIs.

Property functions

ARQ [24] is a query engine for Jena [25] that implements SPARQL. ARQ provides *property functions* as a way to extending ARQ while remaining within the SPARQL syntax, and, with well-written property functions that maintain the declarative nature of a relationship of subject and object named by IRI, new capabilities can be added by applications for local needs without needing to modify the ARQ query engine code base nor the syntax of SPARQL.

A property function causes some custom code to be executed instead of the usual matching of a triple pattern against the graph data. A well-written property function should express declaratively the relationship between a subject and an object although computational restrictions may be necessary. ARQ executes any property functions in a way that respects the position of the property function in the containing BGP so which variables are already bound at that point in the query does not change.

Free Text Searches

The property function mechanism has been used to provide access to different indexing technologies, including Lucene, Autonomy Enterprise Search, Google, Wikipedia. ARQ itself does not provide the free text indexing but provides the bridge between a SPARQL query and the index. The property function implementing the search directives takes a search string accepts and other parameters for controlling the search and return values that are RDF terms. This is usually the URI of the document. In this case the indexing technologies use the body of some arbitrary document as the indexing text which is not part of the knowledge base itself.

3 Data sources

Our approach to the problems of scale and recall posed by identification uses some common structured and semi-structured data sources both to guide the searches to restrict the scale, and to increase the recall. This use of particular data sources contrasts with the generic but document centric approach followed by the EDRM reference model which is suitable for arbitrary e-discovery requests on arbitrary data sources.

We use some common structured data sources (organizational hierarchy and product catalog)

- The organizational hierarchy provides personal contact information for all people working within HP together with information about the reporting structure and a high-level business area names of the organizational structure.
- The product catalog is used by different content management systems within HP to provide different kinds of product related information ranging from product specification to collections of unstructured documents about products intended for use by sales or marketing. In this example, we will use the product catalog to obtain the product name, product number, product line, and business area.
- The product catalog and organizational hierarchy have common fields allowing connections between people and products to be made. For example, the products business area can be used to identify the high-level organization responsible for a product line.

The semi-structured data sources are important because they provide connections between the structured and the unstructured data sources. They connect structured entities to unstructured text which can be used to characterize topics which can be used to search other unstructured data sources. They connect unstructured text to structured entities which can be joined with other structured data sources. So unstructured text in the semi-structured data source can be searched and the entities of the responsive documents retrieved.

- There are many different content management systems within HP which are used to: generate the external HP web site, organize unstructured sales brochures and support information. They associate the product catalog with many different forms of structured and unstructured data. They are an important data source for e-discovery.
- There are several repositories of technical reports for which author, creation date, abstract structured fields are maintained, some of these technical reports grouped by business area as well as maintaining a record of a report's reviewers.
- The email repositories are very important semi-structured data sources for most e-discovery cases. But for an organization the size of HP they are costly to search without narrowing the search down to particular people and time intervals. They associate people and time intervals to unstructured text and titles which can be searched.
- Patent repositories are semi-structured data sources with structured fields linking people, publications and other patents.

4 An e-discovery example

In a fictitious case HP is alleged to have infringed a patent on the use of *impressive print technology* assigned to *Another Photo Print Company*. HP is required to disclose information relevant to: the development and use of this technology in its products, estimating the likely profit associated with the use of this technology, and showing how sales and marketing made use of the technology. HP is the disclosing party in this e-discovery case, and Another Photo Print Company is the requesting party who initiated the subpoena. HP have their own research and development program for print technology and so whilst complying with the e-discovery request HP are also keen to establish any prior art on the development of such a technology.

The subpoena triggers a duty to disclose and preserve all information relevant to the patent violation case. A team is assembled which is responsible for satisfying the legal request and applying legal holds to preserve relevant data. An identification process is initiated to identify the key witnesses, custodians of data sources, and finding the location of data relating to individuals and organizations. The initial problem is to get a better characterization of the topic, the people, organizations and products relevant to the case.

The patent alleged to have been infringed will be cited in the subpoena together with some related patents and cited publications. These cited documents can be used to provide an initial characterization of the topics relevant to the case (e.g. as sets of keywords and phrases). This obtains an initial characterization of the technical areas related to the impressive print technology. Similarly the authors of the cited documents are identified and provide some an initial set of people to seed our searches.

Our approach is to use the structured and semi-structured data sources to expand and corroborate the people, products and topics related to the case. We show examples of some simple tactics for deriving a set of entities from other entities.

These simple tactics can be composed with others to obtain more complex tactics. There is a need for an e-discovery environment to: manage the sets of entities derived by the used of such tactics; support the composition of complex tactics; and recording the evidence of how they relate to the legal case.

Finding relevant products

The alleged patent infringement is generally concerned with printing, but the subpoena did not identify the products that may have used the impressive print technology. The e-discovery process must eventually identify these products because it needs to assess the potential value that can be attributed to the technology. The subpoena did give some initial seeds as relevant topics. So we use a tactic to find products by searching for web pages on the HP site for product names and numbers co-occurring with terms related to one of these topics

Throughout this paper we will use both parameterized queries to represent such tactics, and the convention that the variable for the parameter is prefixed with a dollar sign.

```
select ?product {
     ?product product_catalog:product_name ?name.
     ?product product_catalog:product_number ?number .
     ?doc ext:GoogleSearch(
          "site:http://h10010.www1.hp.com %s %s %s"
          $n $topic ?name ?number)
    }
```

This tactic uses the Google search engine to perform a search of the HP public web sites dedicated to product sales and support. Alternatively we could perform a search of the internal content management system which generates the content for this web site. The `GoogleSearch` property function calling the Google search engine takes three parameters, a `printf` format string, the maximum number of results to be returned, and a list of parameters for the format string. The format string is used to generate a search string contain both the topic and the product names and product numbers which are retrieved from the product catalog during evaluation of the query. The Google site query requires all terms to be present for a web page to be returned. The documents matching the `GoogleSearch` will be returned in the (ranked) order that Google returns them. Not all searches will return any results and so the search will select products relevant to the topic. This example does not rank the products returned as this would require merging the relevance of scores from distinct searches. Although each search shares the same terms contained in the topic, it also has some different terms because of the product name and product number. This tactic provides a means of obtaining products given a topic.

Expanding the set of products

Once we have identified an initial set of products we can find related products using both the product catalog and the organizational hierarchy.

- The product catalog indicates products that are related through being part of series of products addressing a market. For example there will be a printer targeting the consumer market and others targeting the small business office. Over time there will be a series of products addressing this market. Even within these markets products will address different price points and have different specifications.
- The product catalog also gives information about how the product is made and where it is supported. For example, every product has a product line which is the responsibility of some organization. This allows us to group products using the organizational structure as well as the market structure.

For example in the tactic below: a product line is followed to its organization which in turn is followed to the business unit, then the direction along the hierarchy is reversed to find all the product lines and products produced by this business unit.

```
select ?product {
  $seed_product product_catalog:product_line ?seed_pl.
  ?pl_org hp_ba_hierarchy:product_line ?seed_pl .
  ?pl_org hp_ba_hierarchy:business_unit ?unit .
  ?org hp_ba_hierarchy:partof ?unit .
  ?org hp_ba_hierarchy:product_line ?pl .
  ?product product_catalog:product_line ?pl }
```

Finding relevant people

The subpoena provides an initial characterization of the topics related to the impressive print technology. We now examine a tactic for using a topic to obtain a set of relevant people using a semi-structured data source. We use the HP Labs technical reports repository which we have indexed using both the Autonomy Enterprise search engine and Lucene. Semi-structured data sources, such as the technical reports repository and the content management systems, are very important because they allow the structured and unstructured data sources to be used together.

```
select ?person ?doc ?score {
  (?doc ?score) ext:TechReportsTextSearch($topic
                                          $n $relevance) .
  ?doc TechReports:author ?person
}
```

The text search property function used for searching the technical reports repository again takes three arguments (the search string, the maximum number of results and the minimum relevance score). The search returns the document URL and its relevance score which are both returned by the SPARQL select query because they provides the evidence for the relevance of the person to the topic. The results of this query will be in the ranked order returned by the single text search which gives a meaningful relevance score and ordering because a single search was used.

For example, this tactic for finding relevant people to a topic - returned the groups of (fictious) authors of documents relevant to one of the seed topics

- David Shaken, Neil Arrested, Ant Fame, Iris Retinex
- Daniel Lyon, Ron Glass, Gary Circle
- Neil Arrested, David Shaken, Ace Beach
- Daniel Lyon, Ron Glass
- Ron Glass
- David Shaken, Neil Arrested, Ant Fame, Iris Retinex
- Matt Goat, Kelvin Chemistry
- Peter Wilder
- Ernest House

Two of the authors (Ron Glass and Peter Wilder) were also authors of cited papers and/or related patents cited in the subpoena. This provides further corroboration of the relevance of these people to the case. The relevance of an entity to the case is corroborated when the same entities are returned by distinct search paths. Several people who were authors of these cited papers or patents did not show up in the search because they did not write any technical reports, but they did write lots of patents and so a similar query on a patents database would also find patents related to the topic.

So we can derive a larger set of people who are potentially related to some of the seed topics without using any of the people seeds given by the subpoena. Not all of these are as good as each other, but the emphasis in e-discovery is on performing a search with high recall and will not be discarded at this stage. Some of these seeds were directly derived from the patent and some were corroborated by the searches. i.e. when the same people were derived using different search strategies.

Expanding the set of people

There are several tactics with which we can expand the set of people considered relevant to the case using the structured and unstructured data sources. These are simple tactics that take a person and retrieves a larger set of people using only the structured data sources and not used the hybrid search capability. However these simple tactics can be combined with some other tactics deriving a set of people relevant to a topic which can be used to corroborate or rank the expanded set of people. For example, a simple tactic uses the organizational hierarchy to expand the potential set of people. This exploits the heuristic that people in the same group have similar skills or work on similar products (which is not always true but which is still useful during identification), and so would be likely to possess information relevant to the case.

```
select ?person {
 ?manager hp_org:manages $seed_person .
 ?manager hp_org:manages ?person
}
```

Similarly other tactics use the semi-structured data sources. A simple tactic expands the set of people by retrieving the co-inventors of the patents written by the seed inventor.

```
select ?person {
    ?patent Patents:inventor $inventor .
    ?patent Patents:inventor ?person
}
```

A similar tactic uses the technical report repository to retrieve the co-authors of documents patents written by a given person.

Such tactics can be combined with other tactics which derive people relevant to a topic to obtain stricter queries. For example, the tactic following makes the expansion to all organizational peers of the seed person conditional on the manager having written a relevant patent.

```
select ?person ?doc ?score {
 ?manager hp_org:manages $seed_person .
 ?manager hp_org:manages ?person .

  (?patent ?score) ext:PatentTextSearch($topic
                                        $n $relevance) .
  ?report Patents:inventor ?manager

}
```

Expanding a topic to generate related topics

The technical reports repository can also be used to derive other related topics. The repository has title and abstract fields for each document that are good sources of the keywords used to characterize a topic.

```
select ?title {
  ?doc ext:TechReportsTextSearch($topic,
                           $n,$relevance) .
  ?doc TechReports:title ?title
}
```

We obtained an expanded set of topics (fictious) with one of the initial seed topics.

- “Ink location for Color Halftones”
- “Geometric Screening”
- “Fundamental Characteristics of Halftone Textures”
- “Curved dithering”
- “Multi-pass Error-Diffusion for Color Halftones”
- “Lossless Compression of half-toned Images”
- “Anti-aliasing Methods for Computer Graphics Hardware”
- “Inverse Half-toning for Error Diffusion”
- “n-Simplex gamut mapping”

Some of these topics are inappropriate: instead of being concerned with printing they are concerned with computer graphics, image compression, and some whilst concerned with printing are concerned with color gamut mapping!

Corroboration

When different search strategies are used to retrieve a particular entity and the same (or similar) entities (or concepts) are retrieved to the existing seeds then we have provided additional corroboration of these entities. We have already encountered this when we found that “Ron Glass” was also an inventor of a related patent, a publisher of a cited paper, and the author of a technical report responsive to one of the seed topics.

Similarly we can corroborate the relation between a topic and product by deriving people related to topic and requiring them to be in an organization responsible for a product line. Unfortunately if we perform a topic search of the technical reports repository we will find that every author of a responsive document will occur in Labs which is not responsible for any product line. Either we need a different repository, or we need to use a weaker relation between people and products. For a weaker relation we use the existence of an email message relevant to a topic between the person from labs and someone in the business.

The tactic below takes a topic and a product as arguments, plus some controls for the two searches that are used. One performs a search of the technical reports, but the other performs a parameteric search capability on the email repository which restricts the search to emails between two people. The tactic returns the evidence that the topics are related to the product. This evidence is the document and email that are relevant to the topic, and the people who communicated about the topic – one of whom is in an organization responsible for the product.

```
select ?personA ?personB ?doc ?dscore ?email ?escore{
(?doc ?dscore) ext:techreportsTextSearch($topic,$n1,
                                $relevance1) .
 ?doc techreports:author ?personA .
 Sproduct productmaster:product_line ?product_line .
 ?product_line hp_org:is_in_org ?org .
```

```
   ?personB hp_org:is_in_org ?org .
  (?email ?escore) email:EmailTextSearch(?personA,?personB,
                             $topic,$n2, $relevance2) .
}
```

This evidence could then be analyzed to provide a more sophisticated scoring of the quality of the corroborating link between the topic and the product. We need some means of scoring the corroboration that would take into account either the quality of the relevance of the technical report or the email message to the topic, or the number of relevant communications between the two people. At the moment we just return the evidence.

5 Conclusions

We implemented a hybrid search and query by extending SPARQL using property functions which returned ranked search results. This gives a rich retrieval model allowing text search and query of structured and semi-structured data to be used together. This was used to exploit product and organizational structure to increase recall by finding potentially related people and products.

Unfortunately both product and organizational data sources are constantly changing. For our approach to be most effective there is a need to find the organization and product structure for the particular periods of time relevant to the e-discovery request. This need not be information that is kept during the normal operation of a business. For example, it is common to keep only the current organizational information.

E-discovery legislation only requires an enterprise to disclose information that is held by the enterprise for the normal operation of its business. It does not require enterprises to store additional information. In fact, it is for this reason that proactive records management systems are proposed for e-discovery, as they enforce policies which stipulate that only data with a business purpose should be kept. However maintaining historical organizational and product information to help e-discovery is optional because this is not part of the normal operation of a business. Interestingly the product and organizational structure themselves may not be relevant ESI. They are merely a means of finding relevant ESI and perhaps of understanding its significance.

There are further opportunities for the application of semantic technologies in e-discovery:

- Litigation readiness – where semantic technologies can be used to represent knowledge which helps the finding of ESI about people and products and where to find particular types of data that exist already without keeping additional ESI. Information map of the data sources available in the company.
- Representing the semantics of data sampled during e-discovery

- Describing the provenance of data through the different stages of e-discovery

6 References

1. Rothstein, B.J., Hedges, R.J., Wiggins, E.C.: Managing Discovery of Electronic Information: A Pocket Guide for Judges. Federal Judicial Center, 2007. http://www.fjc.gov/public/pdf.nsf/lookup/eldscpkt.pdf/$file/eldscpkt.pdf
2. *SPARQL Query Language for RDF*, Prud'hommeaux, Eric; Seaborne, Andy; Editors, W3C Recommendation, 15 January 2008.
3. “Federal Rules for Civil Procedure”, 1st December 2008, http://www.uscourts.gov/rules/CV2008.pdf
4. Francis, J.C., Schenkier S.I.: Surviving E-discovery. 2006 Magistrate Workshop, http://www.fjc.gov/public/pdf.nsf/lookup/MagJ0608.ppt/$file/MagJ0608.ppt
5. TREC Legal Track home page: http://trec-legal.umiacs.umd.edu/
6. Electronic Discovery Reference Model. http://www.edrm.net
7. EDRM search guide. EDRM 2009, http://www.edrm.net/files/EDRM-Search-Guide%20v1.14.pdf
8. Example Electronic discovery request http://www.fjc.gov/public/pdf.nsf/lookup/ElecDi13.pdf/$file/ElecDi13.pdf
9. Example preservation order http://www.fjc.gov/public/pdf.nsf/lookup/ElecDi21.pdf/$file/ElecDi21.pdf
10. Example meet and confer form http://www.fjc.gov/public/pdf.nsf/lookup/ElecDi22.rtf/$file/ElecDi22.rtf
11. The Sedona Conference http://www.thesedonaconference.org
12. Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information retrieval, Addison Wesley, ACM Press 1999.
13. Surajit Chaudhuri, Raghu Ramakrishnan, Gerhard Weikum, Integrating DB and IR Technologies: What is the Sound of One Hand Clapping?, Proceedings of the 2005 CIDR Conference.
14. Michael J. Cafarella, Christopher Re, Dan Suiciu, Oren Etzioni, Michael Banko, Structured querying of web text, 3rd biennial conference on innovative data systems research (CIDR) January 7-10, 2007, Asilomar, California, USA
15. Tao Chengm Xifeng Yan, Kevin Chen-Chuan Chang, Supporting entity search: a large-scale prototype search engine, ACM SIGMOD’07, June 12-14, 2007, Beijing, China.
16. Roy Goldman, Jennifer Widom, WSQ/DSQ: A Practical Approach for Combined Querying of Databases and the Web, ACM SIGMOD International Conference on Management of Data in May, 2000. http://www-db.stanford.edu/~royg/wsqdsq.pdf
17. Erik Hatcher, Otis Gospodnetić, and Michael McCandless, Lucene in Action, Manning Publications; December 2004; ISBN 1932394281
18. Sesame, the like operator, http://www.openrdf.org/doc/sesame/users/ch06.html#section-like
19. Virtuoso, bif:contains full text search, http://docs.openlinksw.com/virtuoso/rdfsparqlrulefulltext.html

20. Glitter, textlike and textmatch operators, http://www.openanzo.org/projects/openanzo/wiki/SPARQLExtensions
21. The Sesame LuceneSail: RDF Queries with Full-text Search, NEPOMUK Technical Report 2008-1, http://www.dfki.uni-kl.de/~sauermann/papers/Minack+2008.pdf
22. Agissilaos Andreou , Ontologies and Query expansion, Master of Science, School of Informatics, University of Edinburgh, 2005 http://www.inf.ed.ac.uk/publications/thesis/online/IM050335.pdf
23. Crestani, F. Application of Spreading Activation Techniques in Information Retrieval. Artificial Intelligence Review, 11(6): 453-482, 1997
24. ARQ: Home page: http://jena.sf.net/ARQ
25. Jena: Implementing the semantic web recommendations. J. J. Carroll et al. In proceedings of the 13th international World Wide Web conference (WWW2004).