

# Finding narratives of activities through archival bond in electronically stored information (ESI)

Maria Esteva, Weijia Xu, Jaya Sreevelsan-Nair, Ashwini Athalye, Merwan Hade  
Visualization and Data Analysis Group  
Texas Advanced Computing Center (TACC), University of Texas at Austin

## ***Abstract***

*The motivation for this research is to find the trail of documents that belong to an activity in a unstructured collection of electronically stored information (ESI). We use the concept of archival bond to define the relationships between records belonging to an activity. We have developed and tested a paragraph alignment method inspired by a bioinformatics application to find related documents. To enhance visual understanding of these relationships, we are also developing a graphical user interface (GUI) that displays relationships between documents, their authors, and organizational functions. The use of archival concepts as a framework for this research will ensure that the information recovered is complete and that there are contextual elements to support its interpretation. We propose that this framework and its implementation are appropriate for e-discovery purposes.*

## ***1. Introduction<sup>1</sup>***

E-discovery is the process by which electronic information is recovered and preserved from ESI to use as evidence in a court of law. Documents may provide evidence of specific transactions, communications, and projects, but to fully understand the trail of events that result in any given activity recorded in the documents these need to be interpreted [1]. In turn interpretation is aided by contextual information about the case and the provenance of the documentation involved as evidence. These elements: documents as evidence, contextual information and interpretation, give form to the story that the lawyers present in their depositions. And yet, finding evidence in unstructured ESI, in which content may be incomplete or obscured by lack of organization, and information about the documents' provenance may be buried in the chaos of un-ruled and undocumented systems are some of the challenges of e-discovery. It is at this juncture that an archival perspective is relevant to the e-discovery process.

Archivists conceive of collections as formed by groups of related documents. A fundamental concept in archival theory and practice, known as archival bond, describes relationships between documents as an essential property of documents. Luciana Duranti writes that archival bond, "is expression of the development of the activity in which the document participates..." [2] Documents "are" because of their relationship to other documents in a same collection. While all the documents in a collection are bonded through the collection's structure [3], there are stronger relationships between sub-groups of documents that belong to the same activity. Documents

---

<sup>1</sup> While it is common to encounter that the terms records and documents are used indistinctively, in the archival domain it is most accurate to use the term record than the term document, specifically when referring to records that provide evidence of transactions. However for consistency and agreement between the archival and the computer science domains, we will use the term document throughout this paper. See both definitions at: <http://www.archivists.org/glossary/index.asp>

related through archival bond are fundamental to the meaning of the next related document.<sup>2</sup> [4]

Archival bond is strongly linked to other archival concepts such as evidential value, authenticity, integrity, and context. If in a collection the archival bond is broken due to either missing or not genuine documents, the ability of the documents to provide evidence of the activities of their creators is hampered. In turn, contextual information is essential to complement the trail of documents because it provides insight about the circumstances in which the documents were created, stored, used, and maintained. We suggest that these archival concepts are key to e-discovery processes in ESI because they strive to recover the complete set of records that pertain to an activity as well as the contextual information that supports interpretation. This is essential for reconstructing a conceivable story.

To find all of the related documents as a trail of evidence of an activity in unstructured ESI, we considered these archival concepts as key elements in the design of our research. This project combines qualitative research, computational analysis, and visualization to recover and display related documents in relation to their authors and organizational functions. To develop and test this project we chose a particularly challenging collection of Spanish documents, which are varied in length and topics. The qualitative exploration of the origins and development of a corporate ESI was useful to understand the development of this type of unstructured accretion of documents and to find ways to extract meaningful information from it.

## ***2. Conceptualizing archival bond in unstructured ESI***

In controlled electronic record keeping systems, elements such as indexing, time stamps, encryptions, registries, file-naming conventions, metadata, and general rules to follow are used as manifestations of archival bond. These elements make the collection trustworthy, as it is possible to identify related records and context within the system. This is not the case in an unstructured ESI, with no obvious organization schema and somewhat dubious metadata embedded in the file properties.<sup>3</sup> Examples of unstructured ESI are records of various types that accumulate chaotically in shared directories or are dispersed in different media without an explicit organization system.

To conceptualize archival bond in unstructured ESI we use Duranti's and Guercio's definition of archival bond as "the network of relations between records" [5]. In the absence of the explicit elements of nexus mentioned above, we propose that the relationships be based on the documents' contents referring to a target activity. When looking for evidence of an activity it should be possible to recover all of the related documents that can be found in the ESI. For example, considering a typical corporate project as our target activity, we should obtain the preliminary plans and proposals, financial documentation, memos, communications between the

---

<sup>2</sup> According to the SAA Glossary, "The archival bond places a record in context and gives additional meaning to the record. It includes the relationships between records that relate to a specific transaction (such as an application, a resulting report, and the dispositive record that concludes the transaction), as well as the relationship between the records of preceding and subsequent transactions." See definition of archival bond at [http://www.archivists.org/glossary/term\\_details.asp?DefinitionKey=1620](http://www.archivists.org/glossary/term_details.asp?DefinitionKey=1620)

<sup>3</sup> The stability of file properties such as creation and last modified dates embedded in electronic files vary according to the type of creation software used and the way in which those records were saved, transferred or copied to new media.

involved parties, the board approval of the project and documents related to its development and evaluation. Identifying archival bond as a group of related records gathered or generated as a consequence of a common activity in unstructured ESI is impossible without resorting to forms of computational analysis.

Finding archival bond through computational analysis presents various challenges. In the case of text documents, different measures of similarity are used in the form of clustering and classification methods applied to group documents that treat similar subjects. However, our task is somewhat different as our interest is to find documents related to an activity that may encompass different document types and writing styles, and include various sub-topics. These documents may vary in length and therefore contain varied keyword frequencies particular to the topic of interest. Moreover, keywords related to the topic may vary, depending on the writing style of the different authors involved. In addition, records that treat similar topics but are not directly associated with the activity of interest may introduce noise in the trail. The problems get compounded as the size of the ESI increases.

To tackle these problems we have developed a scoring schema based on cosine similarity between paragraphs of documents. Based on a query document related to the project of interest we return a trail of related documents. We call this method *paragraph alignment*. To visualize these relationships along with contextual information, a graphical interface application (in progress) displays the scoring results as documents ranked from more related to less related and in connection to their authors and organizational functions.

### ***3. Ubiquity and diversity in ESI***

For this research, we chose a case of ESI that belonged to a now-defunct philanthropic institution in Argentina.<sup>4</sup> This was a medium-sized project-oriented institution with a hierarchical organizational structure and distinct functional areas. The ESI is made up of individual directories containing the work documents that each staff member accumulated in the shared directory of the institution's networked server for twelve years (1993 to 2005). During that time, these directories were moved from one server to the next every time the institution upgraded its computing systems. During each of these transfers, the file properties including dates that the documents were last saved and modified did not change. This was important for our research as preserving the date properties allowed identifying and sorting documents per year for our study. With a total of seventeen thousand (17,000) Spanish documents, the number of documents per year ranges from 700 to 3500.<sup>5</sup> In turn, the number of staff members whose documents are in the shared directory ranges from 3 to 16 varying throughout the years depending on hires, layoffs, retirements, and job changes. The number of documents per author per year varies as well, including everything from a minimum of 7 documents to a maximum of 300.

Through qualitative interviews conducted before the organization closed to 70% of the staff members that kept their records in the networked server and systems administrators and IT

---

<sup>4</sup> A copy of the institution's ESI was given to us for research purposes.

<sup>5</sup> As this was a multi-national organization, the directories contained roughly equal number of documents in English, Portuguese, and French. It also contained database records, photographs, spreadsheets and email inboxes. For purposes of this research that deals only with Spanish language, documents were sorted out using file management and language sorter software.

consultants, we learned about the staff members' record-keeping and record-making practices. Documents are of two broad types: official and general. The official documents such as board meeting minutes and agendas, annual reports, and calls for grants were prepared by staff from the different functional areas who participated writing small sections that were later compiled and edited by two staff members. General work documents such as memos, correspondence, lists, budgets, narratives, essays, and progress reports were written individually by staff members and in relation to specific projects. There is also a small amount of personal documents in the directories. Some people kept all the versions of a document that they participated in co-writing, including copies of the small sections. Some documents were used as templates into which new content was pasted, while other parts remained the same across versions.<sup>6</sup> Because the institution had steady lines of support, certain terms and names are repeated many times in reference to different activities.

In the absence of explicit record-keeping rules each staff member kept and deleted documents as they pleased, but through the interviews, we learned that a non-explicit practice was to keep more than to delete. This is important contextual information as we can infer that documentation of both the activities as well as the organization are reasonably complete. And yet, this practice of keeping everything introduces the problem of numerous repetitive versions of documents. Overall, within and across directories there are many repetitions as well as a variety of document types and themes. Organization of information was also idiosyncratic. The lack of uniform file naming conventions or directory structure made it almost impossible for people to find other staff members' documents—and even their own—in the shared directory. The study of the electronic record-keeping system is part of the archival perspective that focuses on understanding the ESI's formation process, which in turn can attest to the document's authenticity and integrity [6]. The data obtained through this study is relevant in better understanding of the structure of the ESI and in designing the computational analysis as well as interpretation of its results.

## ***4. Computation of Archival Bond***

### **4.1 Cosine similarity**

Having conceptualized archival bond in unstructured ESI as the content of different documents that refer to a chosen activity, we started by assessing how to identify related documents with existing computational methods. In data mining and information retrieval, one way to find related documents is by measuring the similarity between them. An established measure of similarity between documents used is the cosine similarity [7]. With this approach, each document in a set is transformed into a vector representation and the totality of the documents in that representation constitutes a vector space model. This enables calculating the similarity between two documents by finding the cosine of the angle between the two vectors.

The vector representation of a document is based on the “bag-of-words” model, in which the documents in a set are treated as an unordered collection of terms. Each dimension in the vector is represented by the frequency with which a unique term is repeated in one document and in

---

<sup>6</sup> For example, all the calls for grants contained the same general conditions, all the letters started with the same formula, many quarterly reports contained the same project descriptions used for annual reports or board meeting minutes.

relation to all the terms in the set. This is called a term weighting scheme. The most popular term weighting scheme is tf-idf. There are two components in this schema, the term frequency (tf) and the inverse document frequency (idf). The term frequency is the frequency at which the term occurs in a document, and the inverse document frequency is defined as inverse of the fraction of documents containing the term. The tf-idf weight for each term in every document in the set is computed as the product of term frequency and the inverse document frequency. Thus this weighing function facilitates terms that occur very frequently to weigh less than less frequent terms and hence, could be unique to identify key concepts or topics in the set.

Since the vector space model relies heavily on the term frequency, words such as prepositions and pronouns that occur frequently in any written document are filtered out at the beginning of the computational process. Those words are referred as stop-words. The step of filtering stop-words and mapping the documents into a bag-of-words model is known as document tokenization.

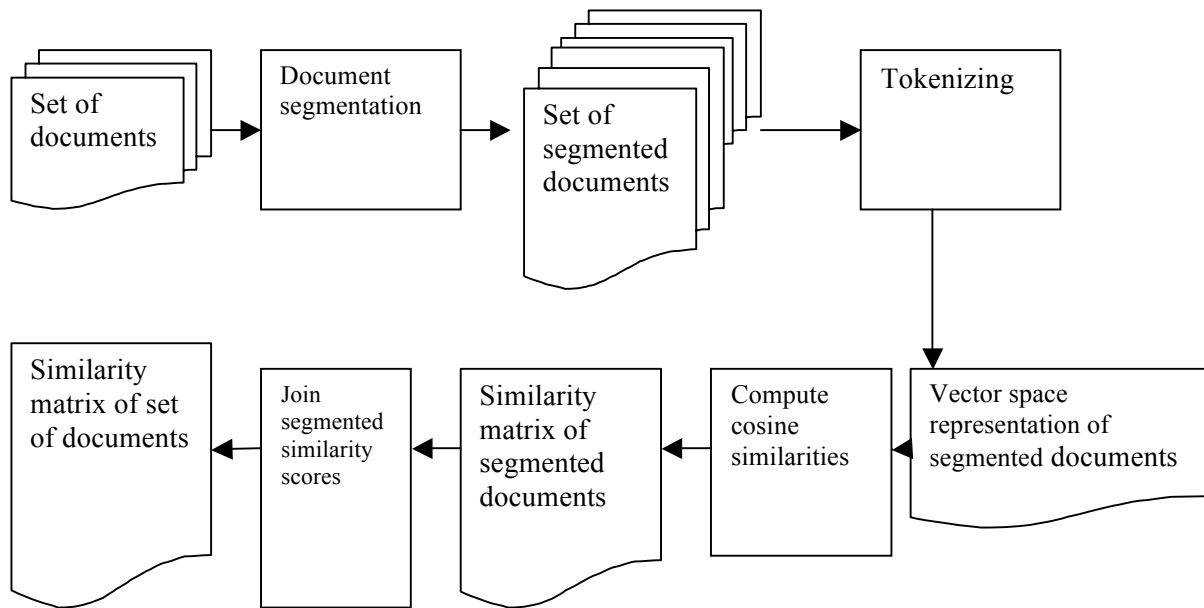
Using cosine similarity for identifying archival bond has specific limitations. Cosine similarity scores quantify the similarity between two documents based on the term frequencies in documents in its entirety. However, in certain cases documents belonging to the same activity may contain multiple topics and therefore may not share a high cosine similarity score with the query document. This is also true for long documents in relationship to short ones that may address the same topic, but due to the substantial difference in document lengths they end up having a low cosine similarity score. To address these issues, we developed a method that computes similarities between documents based on the cosine similarity between document segments.

## 4.2 Paragraph alignment

We draw from local sequence similarity computation, a method used in bioinformatics applications [8]. Biological sequences are represented as one-dimensional strings. While sequences evolve throughout history owing to constant mutation events, the crucial part of the sequences—the part that directly participates in cellular activities—remains relatively stable. Therefore, global similarity between two sequences is often less relevant than the local similarity, which is defined by the highest similarity between substrings of two sequences. Efficient methods for computing sequence similarities often follow a framework in which sequences are broken into n-gram for similarity computations and then assembled to derive an overall similarity [9]. Here we adapt a similar approach which we shall refer to as *paragraph alignment* to determine archival bond between documents that relate to a similar activity. This method is based upon the characteristics of the ESI at hand, in which there are paragraphs repeated across different short and long documents. In the following sections we explain the method's workflow, testing, and its results.

### 4.3 Workflow

Figure 1 below shows our workflow to compute archival bond.



**Figure 1. Workflow to compute archival bond between documents.**

### 4.4 Case study

We tested our methods in the set of documents from the year 1997, with 7 authors and 714 documents. This smaller set illustrates the problems present throughout this particular ESI. The documents in the set have diverse lengths. In general, memos and letters are short documents containing 400 to 600 words, but the board meeting minutes, the calls for grants, and grant forms, may have 4000 words or more. In terms of content, the latter are more diverse as they contain brief accounts about many activities and include general administrative information. Instead, short documents, alternatively, are focused on one activity. In this set, many people and places mentioned for one activity are also mentioned in documents throughout the set in relation to other programs, projects, and partnerships that took place that year. For example, advanced musical training was one of the institution's funding areas; therefore, terms and names associated with the topic are ubiquitous throughout the directories.

To prepare the research set we extracted copies of the documents of the year 1997 from the ESI. For this, all the documents from a given author were sorted by year according to their last modified date, renamed with a file naming convention that allowed keeping track of authors and years, and transformed to ASCII text. We considered that breaking the ESI into yearly sets made sense considering that since in Argentina the fiscal year runs from December to December, an annual set is a self contained accumulation of transactions and activities.

To compute similarity between documents we first break each document into one or more segments. In this experience, each document is segmented based on two properties, a) the paragraphs of the document and b) the number of characters in each document segment. If a paragraph contains less than 1,000 characters, it will be merged into its neighboring segment, i.e,

the preceding or subsequent paragraph. The 1997 set which has 714 whole documents; on segmentation, contains 4,236 documents. This new set consisting of segments of whole documents is used for subsequent processing.

We used Rainbow [10], an open source classification tool, to tokenize the set of document segments and create a matrix of absolute term frequencies. Rainbow was modified to tokenize Spanish diacritics as specified by the ISO/IEC 8859-1 Latin alphabet standard. We used custom-made software to calculate tf-idf and cosine similarity between every other document segment [11]. This computation results in a symmetric matrix in which the values are a similarity score between every other segment. We then processed this matrix to derive the similarity score between the whole documents. In this experience, the score of two whole documents is defined as the maximum similarity score between their segments. The final output is a yearly matrix of cosine similarities between pairwise whole documents which are identified through their file naming convention.

#### **4.4 Comparison between paragraph alignment and cosine similarity**

To evaluate the paragraph alignment method we compared its results with the results obtained by calculating cosine similarity between the whole documents of the 1997 set. The comparison was based on assessing 4 test-groups of selected documents, all of which belong to a target activity. A member of our research team who is familiar with the ESI contents and the institution manually classified the documents. For example, as one of our “test-group activities” we chose to find the trail of documents belonging to a multi-national project in which a renowned master would train young orchestra directors from three Latin American countries in Argentina. The project involved pre- and post- production phases including agreements to accomplish the project, presentation of the project to the institution’s board, coordination between the project managers in three countries, communications with the event site, review of beneficiaries, contacts with orchestras, budgets for the different aspects of the event planning, program marketing, and program reporting. During the manual classification we selected memos, correspondence, appropriation requests, budgets, project summaries, calls for grants, grant forms and schedules, all of which were produced and co-produced by different members in the organization.

Each test-group contained a query document and a set of documents related to the query. The query documents included two memos, a letter, and a form. The related documents included memos, letters and budgets, reports and summaries. They also included calls for grants and board meeting minutes, which are all large documents in which the activity of interest is mentioned along with many other activities that took place in 1997. For each query document, both the cosine similarity and the paragraph alignment methods returned a list of documents ranked from more similar to less similar. Each returned document was labeled as a “true positive hit” if it had been manually classified as related to the query document; otherwise the document was labeled as “false positive hit.” If a document classified as related document was not found among the returned documents it was labeled as “false negative.”

**Table 1. Statistics for each test-group and results returned by each method.**

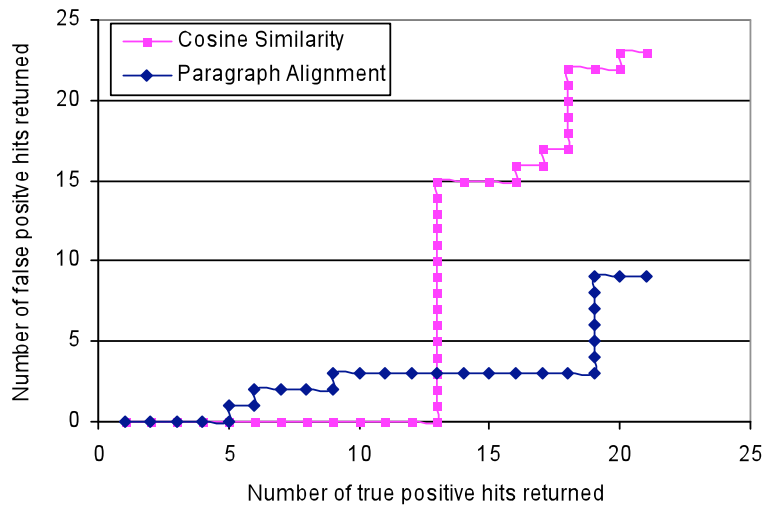
Test Group		1	2	3	4
Number of segments from query document		1	3	5	2
Number of selected documents related to query		21	11	20	3
True positive hits in top 50 documents returned by:	Cosine Similarity	21	8	18	2
	Paragraph Alignment	21	9	19	2

Table 1 shows the number of segments of each query document, the number of manually selected documents, and the number of true positives returned by each method within the top 50 documents. We chose the top 50 documents as our threshold based on our practical experience classifying the test-groups manually. The results show that the paragraph alignment method returns the same number (test-groups 1 and 4) or more true positive hits (test-groups 2 and 3) than the cosine similarity method. Table 2 below shows a comparison of both methods.

**Table 2. Comparison between cosine similarity and paragraph alignment methods**

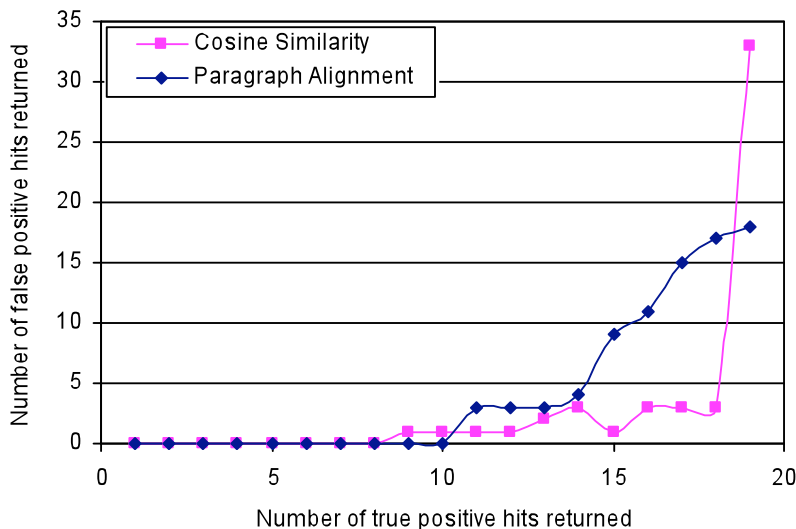
Test Group		1	2	3	4
Number of true positive hits returned		21	9	19	2
Number of false positive hits returned	Cosine Similarity	23	63	33	41
	Paragraph Alignment	9	31	18	5
Sum of rankings of true positive hits returned	Cosine Similarity	376	197	240	44
	Paragraph Alignment	286	117	273	8

For this comparison we considered a list of more than 50 documents returned by the cosine similarity method that includes all the true positives returned within the first 50 ranked documents when using the paragraph alignment method. The row labeled as “number of false positives hits returned” shows the number of false positives returned for each method before the last true positive document returned. It can be observed that there are fewer false positives returned using the paragraph alignment method than the cosine similarity method on entire documents. We also summed the rankings of all true positives. This value indicates how true positives are ranked among all the documents returned. Based on this measure, the paragraph alignment method is less effective than the cosine similarity method only for test-group 3. It should be noted that the query document of test-group 3 is longer than the other query documents (closest to the average length of the documents in the set) and more generic in content (a form). Figure 2 further illustrates the number of false positives found for each true positive.



**Figure 2. Number of true positive found vs. number of false positive found for test-group 1.**

Results show that the cosine similarity method is better than the paragraph alignment method to identify documents with high similarity, such as versions of the query documents. But the paragraph alignment method better identifies related documents with a lower cosine similarity score. Although the latter documents do not share many word distributions with the query document, they are related to the query document by content. For example, they can include a segment or key terms of the query document. In the case of test-group 1 with 21 related documents, the paragraph alignment method returned 10 false positives and the cosine similarity method 22 false positives. Figure 3 below shows similar results for test-group 3.



**Figure 3. Number of false positives found for 19 true positives identified by both methods in test-group 3.**

When we did the query document selection, the criterion we used was that it should discuss the activity of interest, but with no further discussion of how broad or specific the keywords ought to be. What we discovered through this process is that the paragraph alignment method works more

efficiently when the contents of the query document are less generic and more focused on a particular activity.

### 5. Visualization in progress

To display archival bond we are building a graphical user interface (GUI) using C++ with OpenGL [12] and FLTK libraries [13]. The goal of this application is that the entire workflow, from selecting the set of documents to be analyzed to computing archival bond and displaying the trail of related documents, will be achieved by using the GUI as the front-end. As shown in the snapshots in Figure 4, the GUI currently provides the option of displaying the authors as icons, selecting a document as a query, displaying the ranking of related documents up to a threshold selected by the user, and to show the belonging organizational function for each of the documents in a trail.

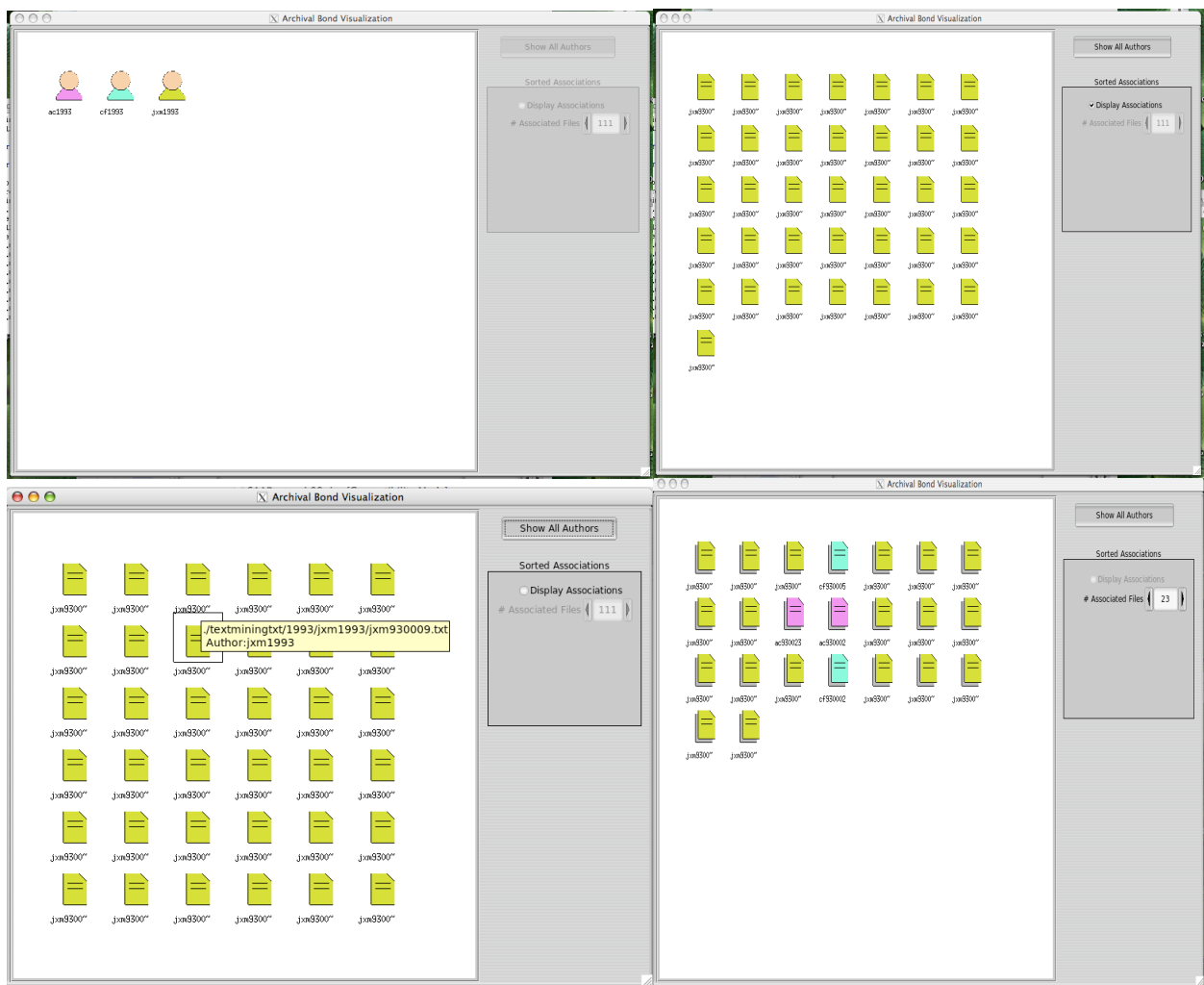


Figure 4. Snapshots of GUI showing clockwise from top left corner a) display of authors, b) display of the records of one author, c) selection of a query record, and d) layout of records showing ranked similarities with the query record.

In Figure 4, each author has a distinct color that is repeated throughout the different displays. In the last screen-shot, showing a group of related documents, authorship can be distinguished by the correspondent color. The visual feature allows determining who worked with whom in an activity and whether other staff members participated. The possibility of analyzing contextual information along with the paragraph alignment results adds evidential value to the discovery process. In the future we will add the possibility to visualize other attributes such as the document's embedded dates as well as the classification of the authors based on functions, roles, gender, etc. The GUI will also provide the option to display the content of the documents. To better understand the roles of the different authors in writing the different parts of a document, we will create visual highlights to identify the parts or versions written by each person. We will further improvise the GUI to include salient features of our paragraph alignment method.

## **6. Conclusions**

In this research we use archival concepts as a framework to find narratives of activities in unstructured ESI through computational methods. We introduce the idea that archival bond can be used to identify groups of records related to an activity, and highlight the importance of context to evaluate the results. We present a novel approach, namely paragraph alignment method, to compute archival bond. This method was designed taking into consideration specific characteristics of a particularly challenging ESI case. The results show improvement over conventional methods of establishing similarity among documents. In this case, the global similarity between documents is less important, as the related documents may only contain portions that overlap the topics involved in the activity of interest. Therefore, it is the “local similarity” what matters in defining archival bond. The same characteristic is observed in biological sequence analysis that inspired our method. In turn, knowing through qualitative research how the ESI developed over time, the way in which staff members created and accumulated documents, allowed for a more accurate interpretation of the results. Our in-progress visualization allows following the trail of related documents in combination with contextual elements to enhance interpretation. Future work includes improving the scoring system and scaling the method to analyze large sets of documents. We suggest that an archival perspective to finding complete stories of activities from unstructured ESI adds value to the e-discovery process.

## **References**

- [1] Gourevitch, P. (2009). The Abu Ghraib We Cannot See. In Sunday Opinion, The New York Times, 24, May p. 10
- [2] Duranti, L. (1997). The Archival Bond. Archives and Museum Informatics. Springer: Netherlands, p. 213
- [3] McNeil, H. (2000). Trusting Records: Legal, Historical and Diplomatic Perspectives. Springer.
- [4] Pearce-Moses, R. A Glossary of Archival Records Terminology. Society of American Archivists, [cited 10 May 2009]. Available at: [http://www.archivists.org/glossary/term\\_details.asp?DefinitionKey=1620](http://www.archivists.org/glossary/term_details.asp?DefinitionKey=1620)
- [5] Duranti, L & Guercio, M. (1997). Research Issues in Archival Bond. Electronic Records Meeting, Pittsburg PA, May 29, 1997, [cited 10 May 2009] Available at: <http://www.archimuse.com/erecs97/s1-ld-mg.HTM>

- [6] Esteva, M. (2008). Formation Process and Cultural Digital Material. In *The Aleph in the Archive: Appraisal and Preservation of a Natural Electronic Archive*, p 52 [cited 10 May 2009]. Available at: <http://repositories.tdl.org/tdl/handle/2152/3840?show=full>
- [7] Salton, G. and Buckley, C. (1988). Term-weighting Approaches in Automatic Text Retrieval *Information Processing & Management*, 24(5), pp. 513–523.
- [8] Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press.
- [9] Wu, S. Manber, U., Myers, G. & Miller, W. (1990) An O(NP) Sequence Comparison Algorithm. *Inf. Process. Lett.*, vol. 35, no. 6, pp. 317-323.
- [10] McCallum, A. K. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, Computer software, [cited 10 May 2009]. Available at: <http://www.cs.cmu.edu/~mccallum/bow>
- [11] Bi H. & Esteva, M. Digital Archives Curation (DAC). Computer software developed for the dissertation: Esteva, M. (2008). *The Aleph in the Archive: Appraisal and Preservation of a Natural Electronic Archive* [cited 10 May 2009]. Available at: <http://repositories.tdl.org/tdl/handle/2152/3840?show=full>
- [12] Open GL. High performance Graphics. Computer software, [cited 10 May 2009]. Available at: <http://www.opengl.org/documentation/>
- [13] FLTK, Fast Light Toolkit. Computer software, [cited 10 May 2009]. Available at: <http://www.fltk.org/>

### ***About the authors***

Maria Esteva has a Ph.D in Information Science from the University of Texas at Austin with focus on digital archiving and electronic records analysis. In her research she uses text mining and visualization methods to make sense of unstructured collections of electronic information. Currently she works full time as a researcher and data archivist at the Texas Advanced Computing Center (TACC).

Weijia Xu is a full time researcher at TACC. He received his Ph.D from the Computer Science Department, at the University of Texas at Austin with focus in data management and analysis. Dr Xu has published in scientific database development, efficient proximity search methods for information retrieval, and information visualization. He currently is co-PI on a NIH funded project to develop computational foundations for comparative sequence analysis based on relational database.

Jaya Sreevalsan-Nair has expertise in the field of scientific visualization for structured and unstructured grids, scattered, and multidimensional data in scalar, vector, and tensor fields. Her recent research is directed towards various algebraic transformations that can be applied to any form of data to convert it to manageable format, for applying visualization techniques. She has experience building graphical user interfaces for various visualization applications. Her Ph.D is from the University of California at Davis.

Ashwini Athalye and Merwan Hade are graduate and undergraduate students respectively in the Computer Science Department at the University of Texas at Austin. They are student assistants in TACC's Visualization Laboratory.