DOE FY2007 SBIR
DE-FG02-07ER84709
FINAL REPORT

Automated Data Cleansing in Data Harvesting and Data Migration

Principal Investigator – Mark Martin

Date: March 18, 2009

Period of Performance
6/20/2007 thru 3/18/2009

Submitted By

Information International Associates, Inc.
1055 Commerce Park Drive, Suite 110
P.O. Box 4219
Oak Ridge, TN 37831-4219
865-481-0388 Voice
865-481-0390 Fax
www.iiaweb.com

Table of Contents

1 Executive Summary

1.1 How the Research Adds to the Understanding of the Area Investigated

In the proposal for this project, we noted how the explosion of digitized information available through corporate databases, data stores and online search systems has resulted in the knowledge worker being bombarded by information. Knowledge workers typically spend more than 20-30% of their time seeking and sorting information, only finding the information 50-60% of the time[1]. This information exists as unstructured, semi-structured and structured data. The problem of information overload is compounded by the production of duplicate or near-duplicate information. In addition, near-duplicate items frequently have different origins, creating a situation in which each item may have unique information of value, but their differences are not significant enough to justify maintaining them as separate entities. Effective tools can be provided to eliminate duplicate and near-duplicate information. The proposed approach was to extract unique information from data sets and consolidation that information into a single comprehensive file.

The proposal also noted that significant work has been performed by the Information International Associates, Inc. (IIa) team to identify duplicate items in a corpus of unstructured data. Intuitively, the same concepts and algorithms used to construct that system should also work for structured bibliographic data. Once the team developed the necessary requirements and constructs to port the existing technology, it was determined that those same concepts and algorithms would most likely work for bibliographic data. However, it was also determined that the challenge to port the same concepts to provide measurable, verifiable, highly accurate results in the Department of Energy (DOE) environment would be an extremely difficult task to pursue. Therefore, a different approach to the problem was developed and is detailed in this report. This research has provided some valuable insight into the difference between (1) removing duplicates in a pipeline processing system and (2) removing duplicates from in a batch of static bibliographic information.

1.2 The Technical and Economic Feasibility of the Methods or Techniques Investigated

As the research progressed, a three-step process for DOE bibliographic data was determined to be most feasible. To test that theory, a small subset was selected from the overall corpus of data that includes over 5,000,000 bibliographic records. The selected subset collection is the DOE Patents Database. Due to the high value of information

[1] "Business Portals: Frameworks for the Extended Value Chain." *The 2005 Delphi Report*: *Insight for Business and Technology Leaders*; October 2005. Delphi Group.

contained in each bibliographic record, DOE determined that any process that was developed and implemented must be quality controlled, at least initially, by 100% inspection by a subject matter expert (SME) in the collection under consideration for duplicate removal. Thus a three-step process was developed, as described below.

1. First, key bibliographic identification fields were extracted from the patent collection, formatted into a structured record, and then sorted. A routine was applied to identify potential duplicates by mere comparison of the identification fields.
2. Second, an intelligent formula, the duplicate density algorithm (DDA), was developed and applied to the file of suspected duplicates in a pair-wise comparison. The DDA calculates and assigns a confidence indicator that suspected duplicates are indeed duplicates.
3. Third, DOE SMEs examined the actual bibliographic records to determine the accuracy of the DDA. Initial research indicates that pairs with high confidence indicators can be automatically removed by an automated process without review by an SME, indicating initial success in the technical approach and in the research and development of the DDA. DOE is currently assessing what level of confidence indicator can be used for the automated removal without compromising the integrity of the database.

The technical processes applied in this research clearly work and are feasible from a time and cost perspective. There is a high degree of confidence that the process of identifying non-duplicates is extremely accurate. The payoff is that more than 75% of the duplicate records will not require any SME evaluation, providing significant economic value in the application of the project technology to existing collections. In addition, of the items identified as potential duplicates, the automated process can safely remove a significant portion of those duplicates. However, owing to the high value of the data, it is unlikely that the technical process can ever achieve the automated removal of 100% of the duplicates without SME review.

1.3 How the Project is Otherwise of Benefit to the Public

Researchers and scientists currently accessing the DOE scientific and technical information (STI) databases will be greatly served by receiving results from queries that will be free of duplicated information. Researchers will not waste valuable research time wading through duplicative information. Researchers will also have continuing higher confidence in the information, since it will be of higher quality. Increasing confidence and quality through the removal of excessive duplicates contributes to the integrity and reputation of the collection and the infrastructure that supports it.

From a larger perspective, government agencies, particularly those with requirements managing bibliographic data such as members of the CENDI organization, can benefit greatly from the ability of the processes developed to be calibrated to expand or contract the scope of potential candidates and data sources according to a specific agency's needs. Commercial application will result in software with great utility for private-sector organizations that broker bibliographic databases. Examples include popular commercial information services such as Lexis-Nexis and Dialog and could be extended to services such as Factiva and other specialized web databases.

Some of the project's technology can also be applied to the integration of legacy databases into more modern web-based access. Any system that is migrating and/or integrating legacy, archived, and/or paper-based, data from multiple sources may need to search for duplicate information. The algorithms developed in this project can be used to identify duplicate records for elimination or archival.

Another beneficial application is the potential for embedding the Intellectual Property of this technology into other software systems. For example, most current federated search software engines have no facility for identifying and removing duplicate entries. It is widely recognized that this creates a significant problem for analysts who must sort through large numbers of duplicate or near-duplicate documents to locate information relevant to their search strategy. Examples of these systems include Convera, InXight, Excalibur, Insightful and a host of others.

2 Comparison of Actual Accomplishments with the Goals and Objectives of the Project

2.1 Summary of Project Activities

The activities of the project were centered on four main areas.

1. The team conducted extensive Internet research for technologies and concepts that could be applied to this problem.
2. The team also collaborated with the customer to understand the exact nature of the problem, along with the subtleties of the data and the underlying principles of data integrity. In addition, the team researched the history of the data's development, including the legacy processing systems and data sources.
3. Based on the nature of the problem and knowledge of technologies available, the team developed a strategy and methodology to solve the problem.
4. The team developed a duplicate removal prototype process to test and demonstrate the selected strategy and methodology.

2.1.1 Internet Research

The Internet activities focused on the following areas:

- Hashing algorithms to assist in assigning a numerical value to textual information that will create clustering of like objects.
- Stemming algorithms for data normalization.
- Packing concepts to remove "noise" from formatted data fields to assist in data normalization.

2.1.2 Understanding the Nature of the Data and the Problem

The team also explored different concepts and approaches that would likely result in maximizing the automated identification and removal of duplicate citations in the overall data store. In the beginning, the DOE Office of Scientific Information (OSTI) input processing was a largely manual system that required a great deal of manual effort to categorize, abstract, and create the bibliographic citations. Bibliographic information from other agencies and other countries that exchanged information with DOE came into the system primarily from magnetic tape, but this data still required a lot of manual manipulation to improve the format and quality to DOE standards. In those days, computer networks, particularly the public Internet, either had not been invented or were not of sufficient bandwidth and sophistication to provide online access to the data. As a result, DOE provided a myriad of hard copy publications to various subsets of the technical community. This created a wide variety of bibliographic collections that reflected publications such as the Nuclear Science Abstracts or Energy Research Abstracts, databases such as the Energy Database (EDB), or collections identified from the source of the data, such as the International Nuclear Information System (INIS) collection (i.e., tapes). This was further complicated by DOE frequently receiving the same citation from several different sources without the sophisticated software to remove duplicates in the processing system. For example, DOE had an exchange agreement with Germany for all Energy Research data. However, Germany provided their nuclear-related research to INIS, so DOE received the German nuclear research from two sources in two somewhat different formats.

Over the years, DOE has migrated the processing of Energy Bibliographic Data from all sources to a very sophisticated, almost totally automated system. During this migration, the legacy collections have been archived into a single data store that includes enough information in the citations and metadata to allow the individual collections to maintain their characters and identities. In addition, all the legacy data and current data are now made available to the research community and the general public through a series of web-based search and access systems. In addition, full-text documents corresponding to many of the citations are also available in the context of the web queries.

In this context, DOE's goal is to remove all duplicated citations in the data store (to the extent possible). Much of the effort was spent reviewing different approaches to parsing data collections to determine the most practical method to locate and remove duplicate citations. The initial approach was to look at the entire data store and proceed with duplicate removal at that level. However, it soon became apparent that there were two types of duplicate citations: desirable and undesirable. For example, the Space Power Citations collection was created as a cost recovery effort on behalf of another agency and must maintain its internal identification and integrity. Also, DOE may not include the entire collection as part of the EDB since only some of the citations reflect DOE-funded research and are already in the EDB. These duplicates are considered desirable duplicates. On the other hand, if the same citation comes into a collection from two different sources (e.g., INIS and Germany as discussed above), this would be an undesirable duplicate.

With this knowledge, and after considering other approaches, the team decided that the most efficient approach was to perform the duplicate removal process on a collection-by-collection basis.

2.1.3 Development of Strategy and Methodology

Following the research activities described above, the team developed a strategy and methodology for duplicate removal. Based on the data analysis, a strategy was proposed to rapidly prototype an automated method to (1) process incoming source data, (2) identify duplicates, and (3) assign a confidence value to those duplicates. The resulting prototype would be reviewed and assessed by DOE SMEs to provide feedback to the development process and to help determine the effectiveness of the techniques used in the methodology.

2.1.4 Development of Duplicate Removal Prototype

The team developed a software prototype that normalizes the input bibliographic source data for processing by extracting, packing, restructuring and sorting bibliographic data fields. The prototype also identifies potential duplicates through the comparison of identification fields. At the heart of the prototype, the team developed the DDA, which calculates and assigns a confidence value to each pair of duplicates identified by the data normalization process. After developing and testing the prototype software, the team tested the prototype on the DOE Patents Database, identifying duplicates and assigning DDA values to each duplicate pair. As part of the testing to determine the accuracy and effectiveness of the DDA, DOE SMEs began to examine the resulting DDA values provided by the prototype.

2.2 Original Hypothesis

The original hypothesis considered a two-fold technical challenge:

1. Significant work that has been performed by IIa's team to identify duplicate items in a corpus of unstructured data as documented in a paper by Coppock, Cooper, and Merrell presented at the 2006 Symposium on International Safeguards at the International Atomic Energy Agency[2]. The basic hypothesis was that intuitively, the same concepts and algorithms used in the construction of this system should also work for structured bibliographic data. The challenge is to port the same concepts in a manner that will provide measurable, verifiable, highly accurate results in the DOE environment.
2. We could select and apply a pattern recognition solution customized for the DOE scenario that will automatically combine the unique information in cluster

[2] *Duplicate Management in Mining Open Source Literature for Knowledge and Intelligence*. Presented at the International Atomic Energy Agency Symposium on International Safeguards, October 2006, Vienna. Co-authors Edrick Coppock and Roy Cooper, Information International Associates; Mary Ann Merrill, InRAD, LLC

documents into the master (or "best") record, thus producing high quality results that meet DOE's current bibliographic standards.

2.3 Approaches Used

The team's approach was as follows:

1. Conduct an in-depth survey of existing data stores and understand the nature of the collections subject to duplicate removal.
2. Conduct interviews and surveys with the SMEs for each of the collections to understand the quality of the data and the nature of the duplication.
3. In addition to the technologies outlined in our proposal, conduct a comprehensive Internet search on other technologies that might be applicable to this problem.
4. Based on all the information collected, develop a strategy for the processes to result in the highest likelihood of success for the amount of effort and cost required to execute the strategy.
5. Select the technologies and processes that provided the best match for the strategy developed, and develop a prototype that can be used to test and demonstrate the effectiveness of those technologies.
6. Test the prototype and analyze the results to measure the accuracy and effectiveness of the prototype.

2.3.1 Survey of the Data Store

As outlined above, DOE has developed a data store of all collections in their STI bibliographic citations. These collections reflect the mission of OSTI over the years, and furthermore, they now, represent a corpus of very high quality bibliographic citations that are in a common, well-defined format. The results of our survey defined the major collections as follows:

- ***DOE Patents Database*** - Citations contain bibliographic descriptions of all patents awarded as a result of research funding by DOE and her predecessor agencies.
- ***Nuclear Science Abstracts (NSA)*** - Citations for the reports, journal articles and other publications related to nuclear energy research funded by the Atomic Energy Commission (AEC) (DOE's initial predecessor agency). The primary mission of the AEC was peaceful development and utilization of nuclear energy. . NSA also includes information from the INIS, which collected data from all member countries participating in the peaceful use of atomic energy.
- ***Energy Data Base (EDB)*** – Citations from two sub-collections:
    - The Historical EDB (recovered from the period when the online database was purged every 15 months)
    - The current EDB.

  Citations represent the research funded by DOE as well as a number of foreign countries with whom DOE has developed bilateral agreements for exchange of energy research information.

- ***Reports Holding File (RHF)*** – A collection of minimal legacy records reflecting the reports that OSTI actually holds in hard copy.
- ***Space Power*** – Collection of records that were created on a cost-recovery basis by DOE on behalf of another federal agency.
- ***SO Scanning*** – A collection of records that were created on a cost recovery basis by DOE on behalf of another federal agency.

2.3.2 SME Interviews and Evaluations

For each collection, SMEs were identified and interviewed as to their assessment of the best process to identify duplicate records within their area of expertise. In addition, SMEs were asked to assess the quality of the citations in their collection. In all cases, the SMEs felt that the preferred process was to identify the potential duplicates within their collections rather than across the entire data store. Several reasons were cited, including defining desirable and undesirable duplicates, as discussed above. Furthermore, each SME felt that the quality of the data within their citations was extremely high based on (1) the rigor of the OSTI Report Processing system, (2) the thorough reviews of records conducted at several manual processing stations, and (3) thorough and sophisticated set of computer checks.

2.3.3 Internet Research

The technical team conducted a significant amount of ad hoc queries for existing algorithms or methodologies to determine duplicates in STI bibliographic data, but they were unable to turn up any concepts worthy of pursuit. Attention was turned to text processing and/or normalization techniques that might be applied to the bibliographic duplicate problem. Specifically, the team considered stemming algorithms, hashing algorithms and text packing techniques that might be applicable to this problem. A number of interesting concepts were investigated.

2.3.4 Strategy Development

After intense discussions and debates, the team agreed that there was little to be gained by creating a sophisticated clustering algorithm for duplicates or near-duplicates, in an automated manner, and combining the information into a single record. This was based primarily on gaining a better understanding of the nature of the bibliographic data and realizing that, owing to the extremely high quality of the data, either record accepted in a duplicate determination would be acceptable, and any information lost in discarding a near-duplicate would be inconsequential. Furthermore, the team decided that the nature of highly structured bibliographic data would not require the application of a rigorous algorithm (such as the Secure Hash Algorithm 1 (SHA-1)[3] described in the proposal. In addition, the team also ascertained that there would be little to be gained by applying the

[3] Federal Information Processing Standards Publication 180-1, 1995 April 17 “Announcing the Standard for Secure Hash Standard.”

Imatch[4] algorithm also described in the proposal. However, the team did believe that the concept of taking normalized “slices” of data from the bibliographic records and building a duplicate assessment methodology did have merit and would be highly effective in the solution of the problem. The concept of normalized data slices is also described in the proposal.

Given these decisions, the team quickly determined that porting the total unstructured duplicate checking methodology to the structured bibliographic world would not result in an acceptable payback. Thus the general strategy was to treat each collection separately in the initial determination of potential duplicates and to use data normalization techniques to assist in duplicate identification and verification.

2.3.5 Technology and Process Selection, and Prototype Development

Given the strategic direction and the results of the research, the team decided to use some data normalization techniques, such as data packing and stemming algorithms on key bibliographic fields to select the candidates for duplicates in each collection. The stemming algorithm selected was the Porter method. This was expected to eliminate 75-80% of the records from any processing or human inspection with a very minimal amount of effort. Once this exercise was completed, the team developed a more sophisticated algorithm using a hash calculation to assign a numerical value to the likelihood of any pair of records identified in the earlier exercise as being duplicates.

The team put these techniques together in a prototype that processed source bibliographic data, identified potential duplicates, and then applied the DDA to these duplicates to provide a measure of confidence that could be verified by SMEs.

2.3.6 Prototype Testing and Analysis of Results to Determine Prototype Accuracy

The team tested the prototype on the DOE Patents Database bibliographic records. The results were examined by DOE SMEs to determine the effectiveness of the prototype processes. Initial test results indicate that over 75% of the duplicates identified by the prototype over a certain confidence value have been verified by DOE SME testers as being undesirable duplicates and can be automatically selected for elimination. This testing review confirmed the accuracy and validity of the prototype processes and technologies.

[4] Chowdhury, Abdur. “On the Design of Reliable Efficient Information Systems.” PhD Dissertation. Department of Computer Science, Illinois Institute of Technology. 2001.

3 Problems Encountered, and Departure from Planned Methodology and Assessment of the Impact on Project Results

3.1 Introduction

Most key technical and design issues have already been discussed earlier in this document. This section summarizes the issues and provides a short explanation of their disposition.

3.2 Issues Identified

- The nature and quality of the data varied from the assumptions made in the proposal
- The necessity of clustering and combining near duplicate records became questionable
- The payback for porting the existing duplicate identification system for unstructured data to work on structured data became questionable
- The utilization of some of the proposed technologies proved inadvisable

3.2.1 The Nature and Quality of the Data

Upon a close examination of the STI bibliographic data store at OSTI, and upon gaining a thorough understanding of the rigor of the legacy processing systems, it became apparent that the quality of the data is extremely high and that every unique record has an extremely high value to the scientific community. The original assumption was that the data had been obtained from multiple sources and loaded into the data stores with little post processing and that this was the source of most of the duplication. On the contrary, all data in the data stores have been subjected to a very rigorous post processing system. The duplication has been introduced by changes in policy, as well as changes in definition of what is to be included in a particular collection. Each record's quality and integrity allow it to stand on its own. While significant time was spent in understanding and qualifying the nature of the source data, the high quality of the data made it easier to identify exact duplicates and also eliminated the need to identify small differences in near-duplicate records, reducing the need for clustering as described below. Therefore, this issue did not adversely impact the progress or results of the project.

3.2.2 Clustering and Combining Near-duplicate Records

The proposal assumes that there are multiple bibliographic citations from multiple sources in each collection in the data store. It also assumes that the data were loaded largely unchecked, resulting in data representing the same technical report but with significant variation in the citations. Had that been the case, it would have been valuable to cluster the citations from the duplicated records, collect each field from the duplicated citations, and select the "best" representation of each field based upon some algorithm.

However, as pointed out above, every record in the data store is of extremely high quality and is acceptable for use as the "authoritative" record. In fact, in the preponderance of the cases, the most variation was in punctuation of a report number or something similar. IIa and DOE concur that little would be gained for this project by performing the clustering exercise. Therefore, the effort planned for researching and developing a clustering component was instead applied to researching algorithms and techniques for determining duplicates in STI bibliographic data (e.g., such as text processing, data normalization, hashing algorithms, and confidence factors). Though this discovery changed part of the planned methodology, it did not impact identification of duplicates, which was the primary project goal.   Payback for Porting Existing Duplicate Identification System

The proposal provides some level of detail about the current duplicate removal process for unstructured data. It explains the roles of the SHA-1, as well as the Imatch algorithm devised by Abdur Chowdhury. The proposal also points out that this algorithm works in the context of a pipeline processing system and that the current system, with its present tuning, is 98% accurate in the removal of duplicated data. While this is extremely accurate for unstructured data obtained primarily within the results from Internet searching, this level of accuracy is not adequate for the static data store of DOE bibliographic data. In fact, DOE's preference is that the data cleansing be statistically near 100% accurate.

After weighing these factors against the likelihood of achieving 100% accuracy by merely porting the present system, and after factoring in the complexity and cost of the technical aspects of the port, it was determined that this approach would not likely be successful. It was decided that a more likely path to success would be to apply the knowledge gained from the processes and techniques developed in the unstructured environment and combine that with the technology used in the process, along with some new concepts.

3.2.3 Utilization of Some of the Proposed Technologies

During this evaluation process, the technologies mentioned above were revisited, including taking a series of normalized 'slices' of terms from a frequency-ordered bag-of-words document and applying the SHA-1 hashing function. This process allows the use of the size, along with the offset of the frequency "slice" that is hashed, to determine the level of discrimination that could applied to determine duplicate items. While this procedure is very effective for unstructured data, we determined that the complexity of this specific process was excessive due to the absence of vagaries in the definition of "duplicate" in bibliographic data. Based on this result, the bag-of-words concept was chosen and implemented with a modified hashing algorithm. We also decided that it would be valuable to normalize the bag-of-words with a stemming algorithm. Though the team did not use all the technologies mentioned in the planned methodology, other proposed technologies were used, in addition to newly identified technologies, , to help achieve the desired project results.

4 Facts, Figures, Analyses and Assumptions used to Support the Conclusions

4.1 Definitions of Accuracy Needed by Duplicate Removal Process

Based on the quality of citations and the high data value, OSTI's goal was for the duplicate removal process be as close to 100% accurate as practical. To that end, the team decided on a three-step process for duplicate removal.

The first step is to ***identify the possible duplicates from a collection within the data store***. For this phase of the process, six bibliographic fields were selected that were most likely to indicate duplicate records. For each field selected, a normalization process was performed by processing the field through either a stemming algorithm or a packing algorithm. A summary of the treatment of each field is detailed in Table 1 below.

Table 1. Methods of Duplicate Checking.

Methods	***Description***
Duplicates by Stemmed Title	Each title is stemmed, stop words are removed, and remaining words are sorted so that each title has less of a chance of being worded differently. The titles are then compared against each other to reveal the duplicates.
Duplicates by Packed and Sorted Report Numbers	For each citation, delimited report numbers are separated, packed and sorted into a single value. Once all citation records' report numbers are built in this consistent manner, only duplicated values are identified for review. Within these duplicates are the analytic records that are improperly marked. These records do not indicate that the parent OSTI_ID or the "IS_Analytic_Flag" field is set incorrectly.
Duplicates by Barcode Number	the legacy_ID value is packed and the results are compared. Only records with matching values in this field are reported as duplicates. Within these duplicates are the analytic records that are improperly marked. These records do not indicate that the parent OSTI_ID or the "IS_Analytic_Flag" field is set incorrectly.
Duplicates by Packed and Sorted Contract Numbers	For each citation, delimited contract numbers are separated, packed and sorted into a single value. Once all citation records' contract numbers are built in this consistent manner, only duplicated values are identified for review. Within these

Table 1. Methods of Duplicate Checking.	
Methods	***Description***
	duplicates are the analytic records that are improperly marked. These records do not indicate that the parent OSTI_ID or the "IS_Analytic_Flag" field is set incorrectly.
Duplicates by Packed EDB Number	The EDB number is extracted from the reference_no field and then packed. Once these special values are compared, only duplicated values are identified for review. Within these duplicates are the analytic records that are improperly marked. These records do not indicate that the parent OSTI_ID or the "IS_Analytic_Flag" field is set incorrectly.
Duplicates by Packed Energy Research Abstract (ERA) Number	The ERA number is also extracted from the reference_no field and packed. Once these special values are compared, only duplicated values are identified for review. Within these duplicates are the analytic records that are improperly marked. These records do not indicate that the parent OSTI_ID or the "IS_Analytic_Flag" field is set incorrectly.

The fields were extracted and processed into records suitable for sorting and examination by a program to identify potential duplicated records. This also accomplishes the task of removing the non-duplicates from further consideration at a confidence level near 100%.

The second step of the process is to ***further characterize the potential duplicates using a modified hashing algorithm*** (the DDA). A summary of that algorithm is in Table 2 below.

Table 2. Duplicate Density Algorithm (DDA).	
Ranking Duplicates	Description
Duplication Density Algorithm	The goal of the DDA is to assign a high degree of confidence (a percentage) that the duplicates reported are truly duplicates and can be quickly removed by using an automated process. As the degree of confidence decreases, a manual review of duplicates is necessary to determine if a citation is to be kept or removed from the database. The density score is based on the number of fields that matched exactly. The higher the density score, the more closely the records match. All the fields that were compared to determine this score are listed as follows:

Table 2. Duplicate Density Algorithm (DDA).

Ranking Duplicates	Description
	• Title (stemmed and sorted) • Report Numbers (packed and sorted) • Barcode (packed) • Product Type • Tile (unaltered) • Author (sorted last names) • Contract Number (packed) • EDB Number (packed) • ERA Number (packed) In detail, the density score is a percentage value of how well the above fields matched when two records are compared. Thus, Density Score = (Total field matches / Total fields with data) * 100.0

The DDA operates pair-wise on potential duplicates and assigns the density score based on the "closeness" of the match of the two records.

The third step in the process is to ***have an SME for that collection to physically inspect the records and verify the duplicates***, observing the density score in that determination. The team's hypothesis is that to identify a density Score above which the duplicated record can automatically be removed. Likewise, the team is working to determine another density score below which we can say the pairs are not duplicated with a high degree of confidence. The team recommends that an SME should examine all records having densities that fall between the high and low scores and make a manual decision.

To date, all DOE bibliographic collections have been through the initial screening process to identify potential duplicates, and all collections have calculated density scores for all marching pairs. This exercise produced the results shown in Table 3.

Table 3. Potential Duplicate Identification Processes.

Methods of Potential Duplicate Identification	Collection Analyzed	Total Potential Duplicates
Records by Stemmed Title	Potential Duplicated Records across All Products	513,554 Records (~256,777 Pairs)
	DOE-Funded Patent Duplicates	8,499 Records (~4,250 ~Pairs)
	Potential Duplicated Records across All Products	187,512 Records (~93,756 Pairs)
	Potential Duplicated Records within DOE Products	137,801 Records (~68,900 Pairs)
	Potential Duplicated Records within OpenNet	22,974 Records (~11,487 Pairs)
	Potential Duplicated Records within Geothermal	0 Records (~0 Pairs)
Records by Barcode Number	Potential Duplicated Records across All Products	56,454 Records (~28,227 Pairs)
	Potential Duplicated Records within DOE Products	56,176 Records (~28,088 Pairs)
	Potential Duplicated Records within OpenNet	84 Records (~42 Pairs)
	Potential Duplicated Records within Geothermal	0 Records (~0 Pairs)
Records by Packed & Sorted Contract Numbers	Potential Duplicated Records across All Products	623,373 Records (~311,685 Pairs)
Records by Packed EDB Number	Potential Duplicated Records across All Products	2,132 Records (~1,066 Pairs)
Records by Packed ERA Number	Potential Duplicated Records across All Products	4,234 Records (~2,117 Pairs)

Once identification of potential duplicates was completed using the methodologies described above, the team proceeded to calculate the density score for all pair-wise potential duplicates. These calculations were performed on two subsets of the data store, the DOE collection and the OpenNet collection. The results of this step are provided in Tables 4 and 5 below.

The Total Comparable OSTI_Ids are the total individual metadata records that formed a matching pair based on one or more methods of duplication checking.

Table 4. Duplicated Records within DOE Products.

Product Group	Total Comparable OSTI_Ids
DOE	2,764,846
Density Score	
100	19,808
89	64
88	616
86	1,886
83	18,450
80	16,244
78	64
75	73,308
71	15,638
67	64,914
63	18,520
60	125,746
57	18,014
56	5,150
50	678,022
44	5,054
43	45,470
40	1,060,926
38	13,358
33	562,756
29	5,514
25	4,054
22	100
20	6,310
17	4,672
14	184
11	4
Density Totals:	2,764,846

Table 5. Duplicated Records within OpenNet.

Product Group	Total Comparable OSTI_Ids
OPN	1,379,814
Density Score	
100	7,202
80	551,284
75	163,064
60	184,136
50	254,640
40	107,684
33	152
25	109,712
20	1,940
Density Totals:	1,379,814

At this writing, DOE SMEs had begun a comparison of these test results with actual source records. From an initial analysis of some of the data, DOE SMEs confirm that all records with a DDA score of 100 were in fact undesirable duplicates. DOE SMEs have also conducted spot verifications to determine that records with a DDA under (50) were indeed not duplicates. While this is only a preliminary analysis, the results have positively confirmed the effectiveness of the prototype approach. Further analysis and testing will provide feedback for refinement of the prototype.

4.2 Assumptions on Needs of Other Applications

Since bibliographic records of STI are largely standardized, particularly across the CENDI community, the processes defined in this report are applicable for other agencies within this community. For similar applications, these processes are easy to modify and should be applicable with a modest amount of effort. By providing an adjustable confidence measure through the DDA, other organizations can modify and control the precision in which records are identified as duplicates to suit the nature of the source data and the needs of the organization.

5 Products Developed under the Award

A large part of the research and work performed in this Phase 1 Small Business Innovation Research (SBIR) project was defined by processes and data manipulation. The key prototype product developed was the DDA. The algorithm has been tested and evaluated in anecdotal instances. However, since DOE is still refining the precise definition of a duplicate record and SME teams are still in the process of being identified to formally evaluate and tune the algorithm, extensive empirical testing has not been performed. However, as noted above, the preliminary evaluations performed by the SMEs have shown that the processes developed do indeed work and appear to be highly accurate. The follow-on work will center on empirical testing, tuning of the algorithm, and identification of thresholds that will allow automated duplicate identification and removal. This work will be fully described in the Phase 2 proposal.

Appendix A: Acronyms and Definitions

Acronym/Term	Definition
AEC	Atomic Energy Commission
bag-of-words	The bag-of-words model represents text as an unordered collection of words, disregarding grammar and word order.[5]
CENDI	Originally stood for Commerce, Energy, NASA, NLM, Defense and Interior; now an interagency working group of senior STI managers from 13 U.S. federal agencies.[6]
data packing	To arrange and align the contents of a data structure (or field) for consistency and processing.[7]
DDA	duplicate density algorithm
DOE	United States Department of Energy
EDB	Energy Data Base, also known as the Energy Science and Technology Database; a file containing worldwide references to multidisciplinary basic and applied scientific and technical research literature.[8]
ERA	Energy Research Abstract
geothermal	A set of geothermal technical and programmatic documents searchable from the web portal http://www.osti.gov/geothermal/; also known as the Geothermal Legacy Collection.
hashing algorithm (hash function)	A mathematical formula for converting data into a representative integer such as a hash sum, value, code, or a simply hash; used for data comparison tasks, to accelerate table lookup, to detect duplicate or similar records, and in cryptography.[9]
IIa	Information International Associates, Inc., the contractor awarded this SBIR.
Imatch (Match Image Management)	A shareware digital asset management application for Windows.[10]
INIS	International Nuclear Information System
NIST	National Institute of Standards and Technology
NSA	National Security Administration
NSA	Nuclear Science Abstract

[5] http://en.wikipedia.org/wiki/
[6] http://www.cendi.gov/
[7] http://en.wikipedia.org/wiki/
[8] http://grc.ntis.gov/energy.htm
[9] http://en.wikipedia.org/wiki/
[10] Ibid.

Acronym/Term	Definition
OpenNet	A web site supported by the DOE's Office of Classification to provide easy, timely access to recently declassified documents and other related information in support of the national Openness Initiative.
OSTI	Office of Scientific and Technical Information
RHF	reports holding file
SBIR	Small Business Innovation Research Program
SHA-1	Secure Hash Algorithm 1, a cryptographic hash function designed by the NSA and published by the NIST as a U.S. Federal Information Processing Standard which computes a fixed-length digital representation (known as a message digest) of an input data sequence (the message) of any length.[11]
SME	subject matter expert
STI	scientific and technical information
stemming algorithm	A method to reduce written words to their base, stem, or root form.[12]

[11] Ibid.

[12] Ibid.