ABSTRACT
We introduce a novel document clustering approach that overcomes those problems by combining a semantic-based bipartite graph representation and a mutual refinement strategy. The primary contributions of this paper are the following. First, we introduce a new representation of documents using a bipartite graph between documents and co-occurrence concepts in the documents. Second, we show how to enhance clustering quality by applying the mutual refinement strategy to the initial clustering results. Third, through the experiments on MEDLINE documents, we show that our integrated method significantly enhances cluster quality and clustering reliability compared to existing clustering methods. Our approach improves on the average 29.5% cluster quality and 26.3% clustering reliability, in terms of misclassification index, over Bisecting K-means with the best parameters.

Categories and Subject Descriptors
H.3.3 [INFORMATION STORAGE AND RETRIEVAL]: Information Search and Retrieval – Clustering.

General Terms
Algorithms, Experimentation.

Keywords
Document Clustering, Ontology, Bipartite Graph Representation, Mutual Refinement Strategy

1. INTRODUCTION
Document clustering was initially investigated for improving information retrieval (IR) performance because similar documents grouped by document clustering tend to be relevant to the same user queries [20] [21]. Document clustering, however, has not been widely used in IR systems [7] because document clustering algorithms were too slow or infeasible for very large document sets in the early days. As faster clustering algorithms have been introduced, they have been adopted in document clustering. Document clustering has been recently used to facilitate the nearest-neighbor search [3], to support an interactive document browsing paradigm [7] [10] [26] and to construct hierarchical topic structures [14]. Thus, as information grows exponentially, document clustering plays an important role for IR and text mining.

In this paper, we introduce a novel document clustering approach that overcomes those problems by introducing a novel semantic-based bipartite graph representation method for documents. Instead of depending on the traditional vector space model, our method models a set of documents as a bipartite graph, where the two sets on the graph are documents and corpus-level significant semantic features that are the selected co-occurrence concepts based on Mutual Information.

This bipartite graph has been used in text mining field such as [18] and [28]. Our bipartite graph representation method is different from the traditional methods in that we use co-occurrence concepts as a bipartite graph node while [18] and [27] used words and terms, respectively. The use of co-occurrence concepts has several advantages. First, the co-occurrence concepts have been regarded as more important than the single word/term [11] [13] [17] because they capture potential relationships between two co-occurring concepts in text. Second, the use of co-occurrence concepts prevents noise terms, which are unrelated to the topic of a document but are found in the document, from affecting the similarity measures during document clustering (to be discussed in more detail in Section 3.2).

After selecting co-occurrence concepts as significant semantic features, they are classified according to their relationships with documents and their semantic similarities in a concept hierarchy in an ontology. Then the documents are clustered based on each document’s contribution to each significant semantic feature group. To refine the initial document clustering, we developed a new spectral co-clustering algorithm that uses the mutual-refinement relationship between the significant semantic feature groups and the document groups so that the two groups are mutually recursively refined.

The rest of the paper is organized as follows. Section 2 surveys the related work. In Section 3, we propose a novel graph-based document clustering method that integrates bipartite graph representation of documents and the mutual refinement strategy. An extensive experimental evaluation on MEDLINE articles is
conducted and the results are reported in Section 4. Section 5 concludes our paper.

2. THE PROPOSED APPROACH: COBRA
In this section, we present our novel clustering method that integrates a bipartite graph representation of documents and the mutual refinement strategy. We call our method COBRA (Clustering Ontology-enriched Bipartite Graph Representation with Mutual Refinement Strategy). COBRA consists of the following three main steps: (1) representing documents as a bipartite graph between the documents and co-occurrence concepts in the documents, (2) initial clustering by grouping co-occurrence concepts, and (3) applying the mutual refinement strategy to the initial clustering results.

2.1 Bipartite Graphical Representation for Documents through Concept Mapping
The first step of all document clustering methods is to convert documents into a proper format. We recognize documents as a set of concepts that have their complex internal semantic relationships. We assume that documents could be clustered based on the significant semantic features (i.e., co-occurrence concepts) in the documents. Therefore, we represent a set of documents as a bipartite graph to disclose the relationships between the documents and co-occurrence concepts among the documents.

The complete procedure of constructing a bipartite graph from a set of documents requires the following three steps: (1) the concept mapping in documents, (2) the selection of corpus-level co-occurrence concepts as significant semantic features of the documents, and (3) the construction of a bipartite graph representation with significant semantic features. First, the concept mapping matches the terms in each document to the Entry terms in MeSH and then maps the selected Entry terms into MeSH Descriptors. We now explain the process. Instead of searching all Entry terms in the MeSH against each document, we select 1 to 3-gram words as the candidates of MeSH Entry terms after removing all stop words from each document. We select those candidate terms that only match with MeSH Entry terms. We then replace those semantically similar Entry terms with the Descriptor term to remove synonyms. We next filter out some MeSH Descriptors that are too general (e.g., HUMAN, WOMEN or MEN) or too common in MEDLINE articles (e.g., ENGLISH ABSTRACT or DOUBLE-BLIND METHOD); see [12] for details. We assume that those terms do not have distinguishable power in clustering documents. Hence, we have selected a set of only meaningful corpus-level concepts, in terms of MeSH Descriptors, representing the documents. We call this set Document Concept Set (DCS), where $DCS = \{C_1, C_2, ..., C_n\}$ and $C_i$ is a corpus-level concept.

In the second step, significant semantic features to be used as a basis for clustering are generated from a set of documents. These significant semantic features indicate the semantic components or the intrinsic meanings of the whole document collection. In order to extract significant semantic features from a set of documents, we take advantage of term co-occurrence in documents. Co-occurrence terms have long been used in document retrieval systems to identify indexing terms during query expansion [6] [13]. For example, in the biomedical domain, co-occurrence has been used to capture potential relationships between genes, proteins and drugs in biomedical literature [22]. In this way, we use co-occurrence concepts as significant biomedical semantic features in the biomedical literature that have been regarded as more important than single term [11] [13] [17]. Given DCS, we define a co-occurrence concept, $CC = \{C_i, C_j\}$, where $C_i$ and $C_j$ are two corpus-level concepts in the DCS. A set of co-occurrence concepts for a document set $V_D$ is represented by $V_{CCD} = \{CC_1, CC_2, CC_3, ..., CC_m\}$, where $m$ is the number of corpus-level co-occurrence concepts for $V_D$.

In order to select co-occurrence concepts from many concept pairs, the Mutual Information [8] is employed. According to the information theory, the Mutual Information of two random variables $x$ and $y$ indicates the mutual dependence of $x$ and $y$ by comparing the joint probability of $x$ and $y$ with the probabilities of $x$ and $y$ independently. A higher Mutual Information between $x$ and $y$ means that $x$ (or $y$) is non-randomly associated with $y$ (or $x$). In this way, the Mutual Information has been widely used to identify lexical dependencies [5], e.g., in finding functional genomic clusters in RNA expression data [4] and in extracting features from large text databases [6][22]. The Mutual Information is defined as follows.

$$\text{Mutual Information}(x, y) = \log_2 \frac{P(x, y)}{P(x) \times P(y)} = \log_2 \frac{f(x, y)}{f(x) \times f(y)}$$  \hspace{1cm} (1)

Here, $f(x, y)$ is the co-occurrence count, which is defined as the number of documents that contain both $x$ and $y$ concepts.

Because the Mutual Information may be “unstable” if $f(x, y)$ is very small, we consider only concept pairs with $f(x, y)$ = 0.05*N, where $N$ is the number of documents [5]. Co-occurrence concepts are mirrored as edges on the graph and their co-occurrence counts are used as edge weights.

In addition, the use of co-occurrence concepts prevents the noise concepts, which are unrelated to the topic of a document but are found in the document (e.g. “Cancer” in this paper), from affecting the similarity measure process during document clustering. For example, suppose a document $D_x$ that belongs to a document cluster (say DC1), has the concepts $\{C_1, C_2, C_3, C_4\}$. $C_4$, however, is not relevant to the topic of $D_x$ (as this paper contains the term “Cancer”). In that case, traditional document clustering approaches may assign $D_x$ to the wrong document cluster (say DC2). In that case, traditional document clustering approaches may assign $D_x$ to the wrong document cluster (i.e. DC2) because $D_x$ contains $C_4$ as important concepts of DC2. However, if we consider co-occurrence concepts, the concept pairs with $C_4$ (as co-occurrence term candidates) would not become co-occurrence concepts due to their very low frequencies over documents. Thus, the irrelevant concept $C_4$ may be removed from the term set of $D_x$ and would not lead $D_x$ to the irrelevant document cluster DC2.

In the third step, we construct a bipartite graph. A bipartite graph $G$ for a given set $V_D$ of $n$ documents and a set $V_{CC}$ of corpus-level co-occurrence concepts is represented as $G = (V_D + V_{CC}, E)$, where $E$ indicates the relationships between $V_D$ and $V_{CC}$. Weights can be optionally specified on edges. In that case one should provide a sophisticated weight scheme to measure the contribution of concepts to each document. However, such a weight scheme may not be appropriate especially for small sized documents, such as Medline abstracts. In addition, the scheme requires $|V_D| \times |V_{CC}|$ complexity. Thus, we draw an un-weighted bipartite graph.
2.2 Initial Clustering by Combining Co-Occurrence Concepts

Here, COBRA generates initial clusters by combining co-occurrence concepts. Since similar documents share the same or semantically similar co-occurrence concepts, COBRA combines co-occurrence concepts and then clusters documents based on their similarities to \( k \) co-occurrence concept groups. On combining them there are two ways to measure the similarity between co-occurrence concepts: their semantic similarity within the MeSH concept hierarchy \( \text{sim}\) and their documents coverage similarity \( \text{sim}_{\text{doc}} \), shown in Equations (2) and (3), respectively.

The semantic similarity between two co-occurrence concepts \( CC_i \) and \( CC_j \) in the concept hierarchy \( \text{sim}_\text{CC} \) is the average similarity of four concept pairs (i.e., the product of \( CC_i \) and \( CC_j \)). The \( C^* \) in Equation 2 indicates the set of parent concepts of \( C \) concept in the concept hierarchy. The document coverage similarity \( \text{sim}_{\text{doc}} \) is the overlap rate of document coverage zones by \( CC_i \) and \( CC_j \).

This similarity is based on the information theoretic based measure [16]. Formally, it is defined as the ratio between the amount of information needed to state the commonality of co-occurrence concepts and the information needed to fully describe what the co-occurrence concepts are in terms of the number of relevant documents.

\[
\text{sim}_\text{CC}(CC_i, CC_j) = \frac{\sum_{C^* \in \text{parents}(CC_i, CC_j)} C^* \cap C^*}{|C^*| + |C^*|} 
\]

\[
\text{sim}_{\text{doc}}(CC_i, CC_j) = \frac{|\text{docs}_{CC_i} \cap \text{docs}_{CC_j}|}{|\text{docs}_{CC_i} \cup \text{docs}_{CC_j}|} 
\]

Here, \( \text{docs}_{CC} \) implies a set of documents that contain \( CC \) co-occurrence concept. We integrate the two measures with weights into the Co-occurrence Concept Similarity (CCS). Given two co-occurrence concepts \( CC_i \) and \( CC_j \), the CCS is defined in Equation (4) as follows: \( \lambda \in [0, 1] \) as weights

\[
\text{sim}(CC_i, CC_j) = \lambda \times \text{sim}_\text{CC}(CC_i, CC_j) + (1 - \lambda) \times \text{sim}_{\text{doc}}(CC_i, CC_j) 
\]

2.3 Mutual Refinement Strategy for Document Clustering

Through the procedures discussed above, COBRA generates initial clusters. However, this clustering cannot correct erroneous decisions as in hierarchical clustering methods. In other words, once clustering procedure is performed, the clustering results are never refined further even if the procedures are based on local optimization.

In our method, COBRA “purifies” the initial document clusters by mutually refining \( k \) co-occurrence concept groups and \( k \) document clusters. The basic idea of the mutual refinement strategy for document clustering is stated:

- A co-occurrence concept should be linked to the document cluster to which the co-occurrence concept makes the best contribution.
- A document cluster should be related to co-occurrence concepts that make significant contributions to the document cluster.

For this mutual refinement strategy we draw another bipartite graph from \( k \) document clusters to a set of co-occurrence concepts. Given the graph \( G = (V_D \cup V_{CC}, E) \), \( V_D \) is a set of \( k \) document clusters (\( V_D = \{DC_1, DC_2, \ldots, DC_k\} \)), \( V_{CC} \) is a set of

<table>
<thead>
<tr>
<th>Algorithm: Initial Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> ( k, V_{CC}, V_D )</td>
</tr>
<tr>
<td><strong>Output:</strong> ( V_{CC}, k ) groups of ( V_{CC} )</td>
</tr>
<tr>
<td>Place each object ( CC_i ) of ( V_{CC} ) as its own cluster, creating the list of clusters ( L ) = {( S_1, S_2, \ldots, S_k )}</td>
</tr>
<tr>
<td>For (</td>
</tr>
<tr>
<td>Find the most similar pair of clusters</td>
</tr>
<tr>
<td>(i.e. Find Max (Eq.(5)))</td>
</tr>
<tr>
<td>Merge ( S_i ) and ( S_j ) into ( S_{ij} ) and Remove them from ( L )</td>
</tr>
<tr>
<td>End For</td>
</tr>
<tr>
<td>/* Now, ( L ) has ( k ) co-occurrence concept clusters */</td>
</tr>
<tr>
<td>Assign each document to the cluster that has the largest number of co-occurrences.</td>
</tr>
</tbody>
</table>

**Figure 1. The Initial Clustering Algorithm**

<table>
<thead>
<tr>
<th>Algorithm: Mutual Refinement Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> ( k, ) Documents, ( V_{CC}, V_D )</td>
</tr>
<tr>
<td><strong>Output:</strong> ( k ) &amp; ( V_{CC} ) groups, ( V_D )</td>
</tr>
<tr>
<td>Do //iteration</td>
</tr>
<tr>
<td>For Each ( CC_i \in V_{CC} )</td>
</tr>
<tr>
<td>Select a document cluster (DC) to which the ( CC_i ) makes the most contribution</td>
</tr>
<tr>
<td>(i.e. Max (Eq.(6)))</td>
</tr>
<tr>
<td>End For</td>
</tr>
<tr>
<td>/* Note the For loop polishes the ( k ) co-occurrence concept clusters */</td>
</tr>
<tr>
<td>Assign each document to the cluster that has the largest number of co-occurrences.</td>
</tr>
<tr>
<td>Loop While NoChangeInDocumentClusters</td>
</tr>
</tbody>
</table>

**Figure 2. The Mutual Refinement Strategy Algorithm**
co-occurrence concepts \( (V_{CC} = \{CC_1, CC_2, CC_3, \ldots, CC_m\}) \), and \( E \) is the relationships between \( V_{DC} \) and \( V_{CC} \). We specify weights on edges so that we measure the contribution of co-occurrence concepts to each document cluster. This contribution is defined as the ratio between the amount of information needed to state the co-occurrence concepts in a document cluster and the total information in the document cluster in terms of the number of documents. The above statement is mathematically rendered as

\[
\text{cntrb}(CC_i, DC_i) = \frac{\text{Size}(DC_i^{CC})}{\text{Size}(DC_i)}
\]

Here, \( \text{Size} \) function returns the number of relevant documents, \( \text{docs}_{DC_i}^{CC} \) represents a set of documents with co-occurrence concept \( (CC_i) \) in the document cluster \( (DC_i) \).

After each refinement, using \( k \) new co-occurrence concept groups, each document is reassigned to the proper document cluster in the same way used for generating the initial clusters. This mutual refinement iteration continues until no further changes occur on the document clusters. Figure 2 shows the pseudo-code of the mutual refinement strategy algorithm.

### 3. EXPERIMENTAL EVALUATION

In this section, we present our test corpora, evaluation measures, and the results of performance evaluation.

#### 3.1 Document Sets

In order to measure the performance of COBRA, we conduct experiments on public MEDLINE documents (abstracts). For the extensive experiments, first we collected document sets related to various diseases from MEDLINE. We use “MajorTopic” tag along with the disease-related MeSH terms as queries to MEDLINE. After retrieving the base data sets, we generate various document combinations whose numbers of classes are 2 to 10 by randomly mixing the document sets. The document sets used for generating the combinations are later used as answer keys on the performance measure.

#### 3.2 Evaluation Method

Clustering approaches have been evaluated by comparing clustering output with known classes as answer keys. There have been a number of comparison metrics, such as mutual information metric [23], misclassification index (MI) [27], cluster purity [29], confusion matrix [1], F-measure [15], and Entropy (see [9] for more examples). Among them we use misclassification index (MI), cluster purity, F-measure, and normalized Entropy. See [25] for more discussion on these evaluation metrics.

#### 3.3 Experimental Setting

In order to evaluate our approach we compare its effectiveness with a leading document clustering approach BiSecting K-means as well as traditional K-means, hierarchical clustering algorithms (single-link, complete-link, and average link), and Suffix Tree Clustering (STC). Two recent document clustering studies showed that BiSecting K-means outperforms both hierarchical clustering methods and K-means on various document sets from TREC, Reuters, WebACE, etc, [19] [2]. A recent comparative study showed CLUTO’s vcluster [31] (with default options) outperforms several model-based document clustering algorithms [30]. This clustering program is an implementation of Bisecting K-means with optimization for large datasets. In addition, the program uses the optimized cluster selection method and a special criterion function that leads to the best overall clustering results in the comparison study of a total seven different clustering criterion functions [29]. According to our previous clustering study [24], CLUTO’s vcluster is significantly superior to the original implementation of Bisecting K-means [19] in terms of clustering quality and scalability.

We provided all the clustering algorithms except STC and COBRA with (word*document) matrixes (i.e. vector representation) as input that are generated by doc2mat Perl script [31]. For STC, we input both a word string and a concept string (we detected MeSH Entry terms in each string and replaced them with MeSH Descriptors).

The implementations of STC are based on [26]. We use BiSecting K-means, K-means, and hierarchical clustering algorithms in the CLUTO clustering package [31]. Because BiSecting K-means and K-means may produce different results for each run due to their random initializations, we ran them five times and averaged the clustering evaluation measure values.

In addition to the comparative experimental evaluation, we also evaluated how much using co-occurrence concepts and the mutual refinement strategy affects the overall clustering quality.

#### 3.4 Evaluation Results

##### 3.4.1 Comparison of Nine Document Clustering Approaches

In Table 1, we averaged the clustering evaluation metric values for the overall cluster quality measure and showed the standard deviations for them to indicate how consistent a clustering approach yields document clusters (simply, the reliability of each approach). The standard deviation would be a very important document clustering evaluation factor because document clustering is performed in the circumstance where the information about documents is unknown. Table 2 shows the relative superiority of COBRA to each clustering approach in terms of cluster quality and clustering reliability. From the tables, we notice the following observations:

- **COBRA** outperforms the eight document clustering methods in all the clustering evaluation metrics.
- **COBRA** has a stable clustering performance (see its standard deviations) regardless of test corpora, while Bisecting K-means and K-means do not always show stable clustering performance even if they yield better cluster quality than COBRA for some datasets.
- The cluster selection method of BiSecting K-means (see “Ir” and “LS” in Table 3) that is used to select the cluster to be bisected, does sufficiently affect cluster performance unlike Steinbach et al’s claim [19].
- Hierarchical clustering methods have a serious scalability problem. They were able to run for only the smallest dataset.
- **STC** does not scale well.
- The performance of STC can be considerably improved by the use of concepts (i.e. by replacing MeSH Entry terms in the strings of documents with MeSH Descriptors).
3.4.2 Evaluation of Mutual Refinement Strategy and Use of Co-occurrence Concepts

We evaluate the mutual refinement strategy (MRS) and the use of co-occurrence concepts to check if each of them is able to improve overall clustering quality. For these evaluations we run COBRA with and without MRS process, and on the uses of co-occurrence concepts or just concepts on the test datasets. Table 3 shows the overall improvements of COBRA through MRS and the use of co-occurrence concepts. We notice that MRS greatly improves the performance of COBRA. The main reason for this result is that MRS, as a variant of EM-type procedure, purifies the initial document clusters by mutually refining \( k \) co-occurrence concept groups and \( k \) document clusters. In addition, we observe that the use of co-occurrence concepts significantly improves and considerably affects the overall clustering quality.

4. CONCLUSIONS

In this paper, we introduced a novel document clustering approach that represents a set of documents as a bipartite graph structure and demonstrated how much the mutual refinement strategy and the use of significant semantic features (i.e. co-occurrence concepts) improve the document clustering results.

The primary contributions of this paper are the following. First, we introduced a new representation of documents using a bipartite graph between documents and significant semantic features (i.e. co-occurrence concepts) in the documents. In addition, using co-occurrence concepts allows us to use more semantically enhanced terms than single-word terms and to remove noise terms in document similarity calculation. Second, we showed that using mutual refinement strategy (MRS) significantly enhance clustering quality. As shown in Table 3, MRS improves cluster quality by up to 47.3% in MI. Third, our extensive experiment results indicates that our approach is superior to all the traditional document clustering approaches and yields the most stable clustering results regardless of test corpora. Our approach showed on the average 29.5% cluster quality improvement and 26.3% clustering reliability improvement, in terms of MI, over Bisecting K-means with the best parameters. Fourth, one should notice that only COBRA supplies the summaries of each document cluster through significant semantic features.

<table>
<thead>
<tr>
<th>MI (the smaller, the better in clustering quality)</th>
<th>Entropy (the smaller, the better)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H(S)</td>
<td>H(C)</td>
</tr>
<tr>
<td>Average</td>
<td>0.50</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>N/A</td>
</tr>
</tbody>
</table>

F-measure (the bigger, the better) Purity (the bigger, the better)

H(S): Single-Linkage Hierarchical clustering, H(C): Average-Linkage Hierarchical clustering, H(A): Average-Linkage Hierarchical clustering, STC(c): Suffix Tree Clustering using concepts, STC(w): Suffix Tree Clustering using words, BsK: Bisecting K-means, lr: selecting the largest cluster to be bisected, LS: selecting the cluster (to be bisected) with the least overall similarity

“N/A” indicates that the corresponding clustering approach fails due to its scalability problem.

Table 2: Relative Superiority of COBRA in Terms of Cluster Quality and Clustering Reliability

<table>
<thead>
<tr>
<th>Cluster Quality</th>
<th>Clustering Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI</td>
<td>Entropy</td>
</tr>
<tr>
<td>STC(c)</td>
<td>K-means</td>
</tr>
<tr>
<td>STC(w)</td>
<td>StsK(lr)</td>
</tr>
<tr>
<td>STC(c)</td>
<td>K-means</td>
</tr>
<tr>
<td>STC(w)</td>
<td>StsK(lr)</td>
</tr>
<tr>
<td>STC(c)</td>
<td>K-means</td>
</tr>
<tr>
<td>STC(w)</td>
<td>StsK(lr)</td>
</tr>
<tr>
<td>STC(c)</td>
<td>K-means</td>
</tr>
<tr>
<td>STC(w)</td>
<td>StsK(lr)</td>
</tr>
</tbody>
</table>

Table 3: Overall Improvements of COBRA through Mutual Refinement Strategy (MRS) and the Use of Co-occurrence Concepts

<table>
<thead>
<tr>
<th>Two Performance Improvement Factors in COBRA</th>
<th>Clustering Evaluation Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutual Refinement Strategy</td>
<td>Use of Co-occurrence Concepts</td>
</tr>
<tr>
<td>Used</td>
<td>Used</td>
</tr>
<tr>
<td>NOT Used</td>
<td>Used</td>
</tr>
<tr>
<td>Improvement by MRS</td>
<td>47.3%</td>
</tr>
<tr>
<td>Used</td>
<td>NOT Used</td>
</tr>
<tr>
<td>Improvement by Co-occurrence Concepts</td>
<td>82.4%</td>
</tr>
</tbody>
</table>
5. REFERENCES


[22] Wren, J.D. Extending the mutual information measure to rank inferred literature relationships, BMC Bioinformatics, 5, 2004, 145


[24] Yoo I., Hu X., and Song I.Y., Clustering Ontology-enriched Graph Representation for Biomedical Documents based on Scale-Free Network Theory, accepted in the IEEE Conference on Intelligent Systems, Sept 4-6, 2006


ACKNOWLEDGEMENTS

This research work is supported in part from the NSF Career grant (NSF IIS 0448023), NSF CCF 0514679 and the research grant from PA Dept of Health