PD-AAW-418
ISN=52378

Contract AID/DSAN-C-0207

Final Report
Development of a Thesaurus of Descriptors for the Development Information System

Project 931-0232: Information as a Tool for Development

March 31, 1981

prepared by

Capital Systems Group, Inc.

for the

Office of Development Information and Utilization
U.S. Agency for International Development

BEST AVAILABLE COPY

Executive Summary

This report gives the findings of a study designed to investigate requirements and alternatives for development of thesaurus of descriptors for AID technical and research reports and project planning, experience, and evaluation documents. Part I describes a survey of AID document users, acquisition and evaluation of thesauri developed by several international development organization and quantitative comparison of these thesauri with word lists and classification codes currently used by AID.

The results of the study show that AID has unique requirements for a thesaurus of descriptors that includes not only scientific and technical concepts but also management information terminology. Users also expressed a preference for a highly refined indexing language structure that shows hierarchical and associative relationships. Users also emphasized their requirements for conversion tools that will allow them to convert previously used terms into new preferred terms.

Based on the evaluation of thesauri from other organizations, and quantitative comparison of these thesauri with AID terms, the study team recommends that AID develop its own thesaurus. As thesaurus development is a lengthy, labor-intensive process, it is also recommended that AID use the UNESCO Thesaurus as a conceptual framework for the development of its own thesaurus. This approach should simplify the level of intellectual effort involved and shorten developmental time.

Alternatives for implementing the recommendations of the study are given in Part II and include an outline for a Thesaurus Data Base Alternative and a Simplified Thesaurus Alternative. The Thesaurus Data Base Alternative involves development of a master computer file that retains information on each AID term. Through development of additional software, the masterfile can be manipulated to produce not only the AID thesaurus but concordance and conversion listings as well.

The Simplified Thesaurus Alternative is a straightforward approach to the development of a simple thesaurus that includes alphabetic, classified, and permuted listings.

Both alternatives are compared in terms of usefulness for meeting user requirements, developmental time and cost. The Thesaurus Data Base Alternative is shown to be a more user-oriented approach than the Simplified Thesaurus Alternative, but far more costly in terms of time and cost.

Part I

Summary Statement--Phase One

1. Background

The purpose of this project has been to develop a thesaurus of descriptors that can be used to index and retrieve AID research and development reports as well as project planning, experience, and evaluation documents.

As it has been our objective to develop a thesaurus that would accommodate the unique requirements of the Development Information System, serve a number of information activities for both AID and other international organizations, and draw upon the expertise and resources that have already been invested in several similar thesauri prepared by AID and other organizations, we have developed the work on the project in two major phases.

Phase One, the subject of Part I of this report, has involved completion of a feasibility study to identify user needs, potential uses for a new thesaurus, and existing thesauri that could be modified to suite AID's requirements.

Based on the recommendations presented in Part I, Phase Two of the project will involve actual production of the thesaurus. A proposal for Phase Two will be given in Part II of this report.

2. Procedure

Our approach to Phase One of the project has involved completion of several activities that were designed to make the process of developing a new thesaurus as objective and scientific as possible. Because AID has already developed two separate thesauri for project documents and research reports, and because several international organizations have published thesauri for information processing activities, we wanted to determine:

- the benefits and limitations of the two AID-developed thesauri;
- the extent to which thesauri produced by international organizations could be adopted as a whole or modified to meet AID's needs; and
- the requirements for developing a totally new thesaurus.

To proceed with our approach, we completed the following project activities:

- interviews with representative users to identify user requirements and preferences;
- examination of information systems, products, and services to determine potential users for a comprehensive thesaurus;
- acquisition and examination of available thesauri on international development;
- evaluation (both qualitative and quantitative) of all available thesauri; and
- analysis of all findings and recommendations for Phase II.

Each one of these activities will be discussed in detail in the following sections.

2A. Interviews with Representative Users

As the purpose of the thesaurus will be to serve a variety of users, we interviewed a diversity of personnel. Interviewees were selected from all phases of the information transfer process--abstractors, indexers, reference librarians, analysts, editors, administrators, project evaluators, project planners, project designers, and policymakers. Interviewees included:

Lida Allen	Director, DS/DIU
Maury Brown	Deputy Director, DS/DIU
Dave Donovan	Chief, Information Division, DS/DIU
Asta Teemant	AID Cataloger
Rick Belcher John Jamula Dick Rader	Abstractors/Indexers, Logical Technical Services Group (LTS)
Garland Standrod	Section Head, Editor of ARDA
Natalie Thompson	Acquisitions, ARDA
Susan Weintraub	Librarian, AID Reference Center
Ellen Boissevain	Librarian, Technical Information Center
Brian Goodhart	Former Analyst
Nena Vreeland	Former Analyst
Chris Wolter	Former Analyst
Maxine Redmond	Former Analyst
Maria Vignone	Former Analyst

Karen Weiss	Independent Consultant, Nutrition
Jan Pellegrino	Contracts Officer
Bernice Goldstein	Project Evaluator, Latin America
Mark Laskin John Doster Jim Massey	Project Designers/Planners in Health and Nutrition
Bill Breslin Brij Mathur	Volunteers in Technical Assistance (VITA)
Warren Wiggins	IQC Field Contractor, Nutrition
Cindy Wincek Bill Anderson	Bureau for Program and Policy Coordination, Impact Evaluation, Rural Development Areas
Frank DiMatteo	Director, Engineering Division

Because the purpose of the user survey was to establish user requirements and preferences, our inquiries focused on the following questions:

- how each user group depends on or interacts with project documents and R&D reports to complete job tasks;
- how each user group evaluates currently available thesauri;
- what each user group would consider to be an ideal thesaurus;
- what special requirements exist for an AID thesaurus; and
- what are other potential uses for a thesaurus.

Although users were interviewed separately, they expressed a consensus of opinion regarding their needs for subject coverage and the conceptual structure of an indexing language. They expressed the following requirements:

1. Comprehensive subject coverage related to international development including the following broad subject categories:

   Education and Communication
   Health and Medicine

Home Technologies and Industries
Food and Nutrition
Environment and Conservation
Public Administration
Business
Industry and Labor
Science
Agriculture
Housing and Construction
Transportation
Energy
Sociology of Development

2. Integration of concepts for management functions and policymaking now expressed by the purpose and technical codes into the technical vocabulary. Example: Funding techniques for immunization programs in Latin America.

3. Addition of terms to describe the contents of evaluation documents such as evaluation methodology and techniques. Examples:

   Comparative Study
   Correlation Study
   Double-Blind Study
   Follow-Up Study
   Longitudinal Study
   Pilot Study

4. Comprehensive geographic coverage that is as specific as possible. Wherever possible, geographic areas should be broken down beyond country, region, and province.

5. Extensive lead-in vocabulary that will accommodate agency jargon and "buzzwords." Terms from the Economic and Social Data Bank (ESDB) and Country Development Strategy Statements (CDSS) should be included in the lead-in.

6. Standardization of word forms and elimination of abbreviations currently used in DIS keyword index.

7. Simple but specific precombined descriptors that are used with a system of generic posting. Essential broad concepts should be identified and included so that more specific, narrower, terms can be added later.

8. A logical structure that includes both hierarchical and associative relationships.

9. Various kinds of display for terms such as tree structures, rotated lists, alphabetical listings.

10. Potential for translation into other foreign languages to facilitate international networking.

In addition to the findings outlined above, there was a significant degree of concern expressed about adoption of a new indexing language. Because countless technical reports have already been classified and indexed through the use of the Dewey Decimal System and Library of Congress Subject Headings, conversion to a new system will be a major undertaking. A similar concern about conversion was expressed by those involved with the technical processing of the project documents.

2B. Examination of Systems, Products, and Services to Determine Potential Users

As a part of our interviews, we also asked how a thesaurus could be used to improve existing systems, products, and services. Potential uses for the thesaurus included:

1. purpose and technical terms to replace codes in PBAR and PAIS;

2. controlled terminology for Objectively Verifiable Indicators for TEXT;

3. general and specific subject categories to arrange and index published bibliographies, publications of Economic and Social Data Bank (ESDB) and Country Development Strategy Statements (CDSS), contracts files, Engineering Newsletter, and Development Digest, as well as to develop profiles for the AID Document Distribution System (ADDS); and

4. an internationally accepted and used thesaurus to assist formulation of a network on international development.

2C. Acquisition and Examination of Available Thesauri on International Development

While we were conducting the user survey, we also compiled a listing of thesauri on international development (see Exhibit A). Although we examined a number of subject-specific thesauri, we acquired only the following external sources for use and evaluation:

UNESCO Thesaurus
Macrothesaurus
VITA Thesaurus of Technical Assistance
UNESCO, Office of Statistics, Draft Classification of R & D Activities by Objective
SPINES Thesaurus

We also obtained available classification schedules and word lists used for information processing within AID. These materials included:

Library of Congress Subject Headings
Development Information (DIS) Keyword List
Research and Development Documents (RANDD) Keyword List
Activity Identification Technical Code List
Activity Identification Purpose Code List
Old Technical Code List
Non-thesaurus terms from Directory of Development Resources
AID Document Distribution Profile
Proposed Code Classification from Burch and Davis Report
Data Element Dictionary from the Economic and Social Data Bank
Business Classification Codes
Financial Classification Codes

2D. Evaluation of Available Thesauri

Evaluation of the available thesauri was completed using three methods that provided both qualitative and quantitative data for analysis:

- evaluation based on criteria established in the user survey;
- analysis and evaluation based on Dagobert Soergel's "Outline for the Analysis of Thesauri..."; and
- analysis of the frequency of matched terms in a merge-match listing.

In making our qualitative evaluation, we focused only on two word lists currently used by AID (DIS and RANDD) and the five comprehensive thesauri developed by other organizations.

The first type of evaluation involved using the findings of our user survey as a list of criteria to rank each thesauri. The results of our evaluation, shown in Table 1, indicated

that the SPINES Thesaurus met most of the requirements expressed by users, followed by the Macrothesaurus second, and the UNESCO Thesaurus, third.

"An Outline for the Analysis of Thesauri/Index Languages/ Classification Schemes" (Exhibit B) developed by Dr. Dagobert Soergel was also used as a tool to evaluate each of the same seven thesauri. The results of the evaluation of each source using this tool are briefly described below:

Development Information System (DIS) Key Word List--Used to index project documents for data entry and to search the B-REF and TEXT files online, the DIS Key Word List consists of an alphabetical listing of highly precombined (3-4 concepts) descriptors that are heavily slanted in subject content toward agricultural topics. All descriptors are enumerated, but no conceptual relationships, either hierarchical or associative, are shown. There is no separate notation given for terms. There is also no evident lead-in vocabulary, instructions, or scope notes. Abbreviations are used, but inconsistently, throughout the thesaurus. Although the conceptual structure of the DIS Key Word List is shown in a separate hierarchical listing, no comprehensive list of main classification headings is given. The hierarchical level of terms is indicated through a display of indented dots. The lack of subject coverage and conceptual structure limits the usefulness of this tool.

RANDD Word List--The RANDD Word List is an informal, alphabetical listing of precombined descriptors (2 concepts) that was developed to index the AID R&D documents and to search the online data base. Subject coverage has been developed as a result of indexer requirements rather than careful subject analysis. All descriptors are enumerated, but no notation is given. The word list does not include a classification schedule, conceptual structure, or lead-in vocabulary. There are no scope notes. Because of its lack of the aforementioned features, it is poorly suited to serve the complex and diverse needs of the Development Information System.

UNESCO Thesaurus--The UNESCO Thesaurus is the working tool for the computerized documentation system of UNESCO. It is used for the indexing

and retrieval of all UNESCO documents and publications. Subject coverage is thorough and comprehensive for most areas of interest in international development except for agriculture.

The thesaurus consists of an alphabetic listing of approximately 12,000 preferred terms, a classified thesaurus, tree structures, and a permuted listing of descriptors.

The descriptors are highly precombined (3 concepts) and enumerated. Because the conceptual structure of the thesaurus is polyhierarchical, two or more notations are given for terms where appropriate. Conceptual relationships are defined, and reciprocals for broader terms, narrower terms, related terms and top terms are given consistently throughout. Scope notes and qualifiers are used extensively and are comprehensive. The lead-in vocabulary is used in both the alphabetic and classified listing. Overall, the UNESCO thesaurus is an excellent and thorough tool that could be made optimal through additional terms for agriculture.

Macrothesaurus--The Organization for Economic Cooperation and Development's Macrothesaurus has been developed for universal use in information processing relating to all aspects of economic and social development. The subject areas covered by the thesaurus are very broad, but include most areas of interest to international development.

The thesaurus consists of several parts: an alphabetic listing of preferred terms, subject category listing, descriptor group listing, hierarchical display listing, and KWIC index listing.

Descriptors are moderately precombined (2 concepts) and tend to be very general. No notation is given, although each descriptor does have a reference to its place in the classification scheme. French and Spanish translations are given for each descriptor.

The conceptual structure of the thesaurus is polyhierarchical and conceptual relationships are well defined. Reciprocals are used consistently. There is limited use of scope notes and qualifiers. The lead-in vocabulary is also limited but indicated consistently with reciprocals.

The physical format of this thesaurus is of importance because of its attractive use of bold, medium, and italic typefaces.

The Macrothesaurus is an excellent tool but is perhaps too general in subject coverage to stand by itself as an information processing tool for AID.

SPINES Thesaurus--The UNESCO SPINES Thesaurus was developed for universal use in information processing in the fields of policymaking, application, and management as related to science and technology. Its subject coverage is the most thorough and comprehensive of all with respect to AID's requirements for both management information concepts and scientific/technical vocabulary.

The thesaurus, an example of thesaurus "overkill," consists of an alphabetic, structured list, terminological graphic displays, and a list of top descriptors.

The SPINES Thesaurus contains approximately 7,000 descriptors that have a variable amount of precombination. Core subject areas are represented by highly precombined descriptors (3-4 concepts) and related subject areas have descriptors that are moderately precombined (2 concepts). Descriptors are designated by a term and notation.

The fault of the thesaurus is that it is perhaps too comprehensive and thorough. All relationships are defined, and all reciprocals are consistently listed. The thesaurus includes scope notes, qualifiers, and an extensive lead-in vocabulary. But, in spite of its strengths in these areas, too many conceptual and terminological relationships have been drawn. The terminological displays are so complex they are virtually useless.

In spite of its excellent subject coverage, the conceptual structure of the SPINES Thesaurus would have to be revised extensively and greatly simplified to be of use to AID.

VITA Thesaurus of Technical Assistance--The VITA Thesaurus was developed by Volunteers in Technical Assistance, Inc., for use in indexing and

retrieving both technical report literature and personnel records. Its subject content is quite applicable to AID's requirements for technical and management information terms.

The thesaurus is divided into an alphabetic listing, a classified listing, and a quick reference alphabetic listing of descriptors with their search codes. Special sections of country, language and business affiliation codes complete the volume.

The alphabetic listing comprises the bulk of the thesaurus and includes approximately 2,000 descriptors, lead-in terms, a code notation for each descriptor, scope notes, qualifiers, and a broader term one-level up for each descriptor. No narrower terms are given for descriptors in the alphabetic listing nor are associative relationships shown.

The conceptual structure of the VITA Thesaurus is hierarchial, not polyhierarchial. A summary of the classification system is given at the beginning of the classified listing.

Because of its subject content and approach to subject classification, the VITA Thesaurus has general value to this project. However, inadequacies in the conceptual structure of the Thesaurus limit its usefulness.

Draft Classification of R&D Activities by Objective--The Draft Classification is a provisional document developed by UNESCO's Office of Statistics. It has been designed for use in processing information related to R&D activities and technology transfer. Its subject coverage is outstanding as the classification is an integrated vocabulary of detailed technical and management information terms. Its coverage of administration, planning, and policy of development concepts is unsurpassed by any other thesaurus or listings studied.

Because it is a draft classification, the schedule is quite unwieldy and impractical for direct adoption. It consists of a general classification schedule broken down into eight categories, and then very detailed listings for each of the eight categories. Each category listing is prefaced by lengthy notes that detail the rationale about

the selection and arrangement of terms. The category listings give a polyhierarchial arrangement of concepts with specific references to related terms in other categories. There is no alphabetic listing of descriptors as in other thesauri, so conceptual and terminological relationships are determined through inspection by the user.

The concepts delineated in the classification range from being suitably expressed in one or two words to unwieldy masses of three or four lines of text. Descriptors as such do not exist although some of the shorter terms could serve as descriptors.

Although the subject coverage of the draft classification is well thought out, it is too unrefined to be adopted as it presently exists. Its conceptual structure could provide a worthy addition to an existing tool.

In summary, our evaluation of all thesauri based on Soergel's "Outline..." revealed that the existing thesauri currently used in-house by AID have limitations that impair their usefulness. In addition, we determined that of the other available thesauri, the UNESCO Thesaurus has the best indexing language structure to meet AID's expressed requirements. We also found the Macrothesaurus to be quite acceptable, but not as finely structured for AID's needs as the UNESCO Thesaurus.

Our third method of evaluation involved generation of a "merge-match" listing which compiled quantitative data on the frequency of matching terms among the two AID-developed thesauri, the four comprehensive thesauri (at this point we eliminated the SPINES Thesaurus from further consideration), and the additional classification schemes and word lists used internally in AID. Also included in the data was a list of key words that project staff had extracted from examination of user search requests.

The purpose of this form of "merge-match" evaluation was to determine whether or not any of the four published thesauri had a significant frequency of matching terms with the two AID-developed thesauri and other internal AID vocabularies. We also wanted to determine how well any of the available thesauri or word lists matched actual user language as expressed in search requests.

Although we could obtain the AID-developed thesauri and two of the four comprehensive thesauri in machine-readable form, we were required to undertake a lengthy process of data entry and conversion programming for the other thesauri and word lists.

The results of the "merge-match" evaluation revealed the most direct character for character matches between the set of AID internal word lists and the UNESCO Thesaurus, followed by the Macrothesaurus, Vita Thesaurus, Draft Classification, Birch and Davis Proposed Codes, and User Terminology in descending rank order.

These results are summarized in Table 2. Table 3 shows the percentage of matches for each external source as compared to total terms contained in each source. Tables 4-12 illustrate the frequency of matched terms between each of the individual AID vocabularies and the external sources. Tables 4-15 are arranged in order of the relative importance of the internal source (DIS first, RANDD second, Library of Congress Subject Headings third, and so on).

2E. Analysis of Findings and Recommendations

As a result of our user study and evaluation of available thesauri, the following conclusions are evident:

- no single external thesaurus provides the appropriate blend of technical and management information terms to serve as an alternative for the diverse vocabularies currently used by AID;
- development of a totally new thesaurus uniquely suited to AID's requirements that incorporates most of the 11,000 terms included in all internal vocabularies would be desirable but also costly and time consuming;
- there is an expressed desire to be able to convert the terminology currently used into new thesaurus term equivalents; and
- as a result of the decision to convert all AID internal vocabularies and available external thesauri into machine-readable form, AID has a significant degree of flexibility in developing a new thesaurus.

Based on these findings, it is our recommendation that AID should:

- develop a new thesaurus;
- use a derivative approach to thesaurus building by basing the content on one of the available thesauri; and
- ensure that it will be possible to develop conversion tools if desired at a later date through careful planning.

This approach would ensure that AID would have a good working tool that includes all of the essential terminology and concepts that are unique to its mission and operation in a strong indexing language structure. It also shortens the amount of time for development, as the new thesaurus would be based on an existing source.

Finally, the proposed alternatives provide an efficient method of production through the use of machine-readable data already captured for the "match-merge" evaluation. Further refinement of the machine-readable data could provide conversion tables or concordance listings if desired at a later date.

Our approach and alternatives for development of a new AID thesaurus are detailed in Part II.

Part II

This section of the report details plans for the development of a new AID thesaurus based on an existing, published thesaurus. It is divided into the following parts:

- approach to thesaurus development;
- alternatives for implementing the approach;
- proposed time and cost; and
- proposed project staffing.

1. Approach to Thesaurus Development

Our approach to thesaurus development would be to use all AID thesauri, word lists and vocabularies and to combine them into a coherent conceptual structure through use of the classification system of an existing thesaurus. We envision the final thesaurus to contain 3-4,500 terms.

We advocate a derivative approach that uses an existing thesaurus as a conceptual framework for the AID thesaurus because it is a labor-intensive and lengthy process to develop a thesaurus as an original effort. We believe a total development approach should be avoided for the reasons detailed below.

One of the objectives of this project has been to develop an authoritative thesaurus that would be accepted and used by other organizations. To be accepted by these groups, all parts of a thesaurus developed as an original effort would have to be subjected to rigorous analysis and review. In order to develop 8 to 10 classification schedules of acceptable quality for the AID thesaurus, both subject experts and lexicographers would be required. In addition, another group of subject experts would be required to review the draft schedules several times and to recommend revisions. This one task could require from ten to twelve months for completion. Added to the length of time required would be the labor expense for subject experts and reviewers as well as the administrative cost of coordinating and monitoring a schedule for completion of the task.

A similar project to the AID thesaurus, the Art and Architecture Thesaurus, has been planned as an original effort for the National Endowment for the Humanities (Contact: Dora Crouch, Center for Advanced Study in the Visual Arts, National Gallery of Art, 737-4215, extension 598). The project will require approximately two years for completion; the estimated cost has been given as $403,347.00.

It has also been estimated that the UNESCO Thesaurus required three years for its total development and cost approximately $200,000 per development year.

As detailed in Part I of this report, we found that the UNESCO Thesaurus or the Macrothesaurus meets many of the users' expressed requirements and has an adequate conceptual structure to serve as a skeleton for an AID thesaurus.

The UNESCO Thesaurus is a more valuable tool when evaluation is based on Soergel's "Outline" and the results of the "merge-match" listing. Although its terminology is weak in agriculture, the conceptual structure of the indexing language is highly refined and structured, and there are more character-for-character matches between the UNESCO Thesaurus and AID vocabularies.

The Macrothesaurus receives a higher rank than the UNESCO Thesaurus when evaluation is based on the criteria determined from the user survey alone. It has a good general coverage of all subject areas and the capability of foreign language translation. It received the second highest number of character-for-character matches in the "merge-match" evaluation. Another point for consideration is the fact that AID is actively involved in a document exchange program with the International Development Research Centre (IDRC). The IDRC uses the Macrothesaurus as its indexing tool.

Although it appears to be a fairly even choice between the two sources, it is the recommendation of Dr. Dagobert Soergel, Professor of Library and Information Science, University of Maryland, that the AID thesaurus should be derived from the UNESCO Thesaurus. Any lack of subject coverage for agriculture will be augmented through the excellent and authoritative coverage of the subject in the DIS Keyword List and other AID sources. Also, the more highly developed classification system will be more hospitable than the Macrothesaurus to both the technical and management information terminology that would be essential to include in an AID thesaurus. The UNESCO Thesaurus also incorporates a system of notation that can be adopted for the AID thesaurus. The importance of notation should be emphasized, as it can be used for the physical arrangement of materials or as a code shorthand for descriptors. The Macrothesaurus does not have a system of notation, and one would have to be developed if the Macrothesaurus were to be used as the basis of the AID thesaurus.

2. Alternatives for Implementing the Approach

2A. Thesaurus Data Base Alternative

In addition to using the UNESCO Thesaurus as the conceptual framework for an AID thesaurus, Dr. Soergel also advocates

development of a thesaurus data base as an alternative approach for preparation of the final product. He supports this approach for several reasons:

- all of the AID internal vocabularies are already in machine-readable form;
- AID terms that match UNESCO Thesaurus terms would be sorted into an appropriate classification schedule by machine, thereby shortening the time and cost of this operation;
- data necessary to develop conversion tables and concordances would be maintained in machine-readable form and in a format that would ensure cost-effective generation of these materials; and
- it would also be possible to generate subsets or mini-thesauri of the indexing language for each operational division. These short lists would enhance searching capability.

The development of the thesaurus data base would proceed in the following steps:

1. Creation of the AID Thesaurus Subset.
   All AID terms would be selected from the data base developed for the merge-match evaluation. All user terminology that was extracted from search requests would be included as well. All information regarding the source of the term would be retained. When an AID term matched a UNESCO Thesaurus term, all data from the UNESCO Thesaurus would be transferred to the subset.

   When an AID term matched a Macrothesaurus term, it would be possible to output all of the Macrothesaurus data into a special file for comparison at a later date.

   Through a short program to eliminate plurals and duplicates, the number of terms would be decreased from 11,000 to 9,000 terms. At this point, the subset would be output into some form that would permit efficient manual review--either punch cards or printout. We would also insure that the subset was maintained in a machine-readable form so that extensive data entry would not be necessary.

2. Intellectual Processing.
This step would represent the bulk of the intellectual processing for the thesaurus and would be essentially a manual operation. It would involve assignment of AID terms to UNESCO hierarchies, elimination of synonyms, and development of the lead-in vocabulary. All terminological and conceptual relationships would be established. Some data entry would be necessary to build the hierarchies and establish reciprocals for broad, narrow, and related terms. The content would be reduced to 3-4,500 terms.

3. Development of the Thesaurus Data Base.
This step would involve an INQUIRE programmer to design and develop software to convert and store the AID thesaurus subset into the thesaurus data base. A primary objective in this step would be to retain all information about the AID terms and their former place in one of the old AID thesauri, words lists or vocabularies. Through this step, the data would be maintained in a format that would eventually allow generation of conversion tables, concordance listings or mini-thesauri.

4. Development of AID Thesaurus Printouts.
An INQUIRE programmer would develop programs to generate the alphabetic, classified, and KWIC listings of the thesaurus.

5. Generation of Printout Listings of the AID Thesaurus.
This step would involve production of draft listings for each of the three parts of the thesaurus.

6. Review and Editing.
AID personnel and project staff would review the content of the thesaurus listings for accuracy, consistency, and thoroughness.

7. Revision of Draft Listings.
The draft listings would be revised and corrected through online editing.

8. Production of the Final Volume.
Following revision of the drafts, a final tape would be delivered to AID for photocomposition of the completed thesaurus.

This approach, advocated by Dr. Soergel, incorporates many features that would be of lasting value to AID.

Although the scope of work is clearly beyond the project description outlined in the original contract, the plan presents a logical scheme for meeting current system and user needs and planning for future requirements. Because conversion tables and concordance listings are a part of the initial planning, it will be possible to generate them if desired at a later date in a highly cost-effective and efficient manner.

Although this plan is logical, there are a number of questionable areas regarding its feasibility. It should be emphasized that this is an innovative, developmental approach, and there will be, no doubt, technical difficulties which will delay speedy completion.

We also anticipate that this approach will involve considerable developmental expense.

The expense of this approach, however, should be juxtaposed with the potential cost involved with development of conversion and concordance tools at a later date.

2B. Simplified Thesaurus Alternative

As detailed in the project contract, AID has specified that the purpose of the project is to develop a useful thesaurus of descriptors to assist in information processing and retrieval. Although the following approach could be modified to include development of conversion tools and concordance listings, the steps outlined below have been designed to produce a thesaurus in a simple, straightforward manner:

1. Creation of the AID Thesaurus Subset.
   As described in the thesaurus data base approach, all AID terms and corresponding data from the UNESCO Thesaurus would be transferred from the "merge-match" evaluation data base to a separate file. All user terminology from the "merge-match" data base would be transferred as well as all information regarding the internal source name of the AID terms. The creation of the subset would require assistance from an INQUIRE programmer and would establish a baseline of approximately 11,000 terms to use in developing the final product.

2. Development of UNESCO Skeleton for AID Thesaurus.
   After the data for the thesaurus had been separated from the "merge-match" data base,

another program would identify AID terms that have UNESCO equivalents. The AID thesaurus subset would be divided into two files: AID terms with UNESCO matches and those without corresponding UNESCO terms. When an AID term matched a UNESCO term, all data attached to the UNESCO term would be added to the AID term. The product of this step would be an alphabetic listing for each of the two files.

3. Preliminary Edit of AID Thesaurus Subset.
Both printouts would be given a preliminary editing to eliminate plurals, duplicates, and inverted terms. Essentially a manual operation, this step would reduce the AID thesaurus subset to a working file of about 9,000 terms.

4. Refinement of AID Thesaurus Subset.
Both files would be refined trough a simple "delete" program. All terms eliminated in Step 3 would be deleted from both files. The resulting printouts of both files would be essential to the next step.

5. Manual Intellectual Processing.
As described in the thesaurus data base approach, this step would entail the bulk of effort for the thesaurus development and would be a manual operation. Both files of terms would be examined. All AID terms, those with UNESCO equivalents and those without, would be assigned to hierarchies. Synonyms would be identified and eliminated as required. The lead-in vocabulary would be developed. All terminological and conceptual relationships would be established. Because many AID terms would have no UNESCO equivalent, this step would require careful analysis and design of a work plan at the beginning so that all data for the new thesaurus would be captured with a minimum of iterations. The subset would be reduced to approximately 3-4,500 terms, and all information regarding the initial AID internal source of each term would be eliminated.

(As this step is the first place where any data necessary for a conversion or concordance listing would be destroyed, AID may want to elect to develop two complete printouts of the AID thesaurus subset. One set of data would be a working set. The other printout

would be an archival file that would be marked to show all decisions regarding assignment to hierarchy, elimination of synonyms, and so on. As this possibility would retain data on all AID terms, it would also add an additional cost in terms of time and labor to the project not calculated for this alternative.)

6. Revision of AID Thesaurus Subset.
The results of Step 5 would be converted to machine-readable form. A service bureau would be employed to ensure fast, efficient data conversion.

7. Development of AID Thesaurus Printouts.
Programming assistance would be required to develop programs to generate alphabetic, classified, and permuted listings of the thesaurus.

8. Generation of Printout Listings of the AID Thesaurus.
This step would involve production of the draft listings for each of the three parts of the thesaurus.

9. Review and Editing.
AID personnel and project staff would review the content of the thesaurus listings for accuracy, consistency, and thoroughness.

10. Revision of Draft Listings.
The draft listings would be revised and corrected as indicated.

11. Production of Final Thesaurus.
Following revision, a final data tape would be delivered to AID to use in photocomposition of the completed thesaurus.

2C. Comparison of Alternatives

Both approaches can be compared in terms of

- the extent to which the final product meets expressed user requirements;
- the length of time required for completion; and
- cost for development.

The Thesaurus Data Base Alternative provides:

- a thesaurus that is optimally responsive in terms of meeting user requirements;
- a data base that maintains many options for the development of additional products to assist in indexing and searching and that may be the most cost-effective approach in the long run; and
- an opportunity to participate in an innovative, experimental approach to thesaurus development as well as to gain appropriate recognition for the results.

It will require:

- a considerable allocation of resources for development;
- tolerance for delays and technical difficulties; and
- extensive INQUIRE programmer support.

The Simplified Thesaurus Alternative:

- meets project expectations as specified in the original contract and will provide a thoroughly satisfactory product;
- can be completed in a shorter time than the data base approach;
- requires a minimum of data processing support; and
- will be a very cost-effective method to obtain a thesaurus.

In contrast to the Thesaurus Data Base Alternative it will not:

- provide flexibility for the generation of other products at a later date; or
- involve developmental methodology that will provide professional recognition for AID's information activities.

3. Proposed Time and Cost

In our opinion, developing a thesaurus in accordance with the Thesaurus Data Base Alternative will require approximately 15 months of developmental time and will cost an estimated $106,000.

The Simplified Thesaurus Alternative can be completed in six to eight months at an approximate cost of $50,000.

Exhibits and Tables

EXHIBIT A

Bibliography of Recent Thesauri Related to International Development

A. Comprehensive Thesauri

Aitchison, J., comp. UNESCO Thesaurus. Paris: Unesco, 1977, 2v.

UNESCO Secretariat, Spines Thesaurus: A Controlled and Structured Vocabulary of Science, and Technology for Policy-Making Management and Development. Paris: UNESCO Press, 1976, 4v.

Viet, J., ed. Macrothesaurus for Information Processing in the Field of Economic and Social Development. Paris: Organization for Economic Cooperation and Development, 1978.

Viet, J., ed. Thesaurus of Industrial Development Terms. Vienna: UNIDO, 1976.

Volunteers in Technical Assistance. VITA Thesaurus of Appropriate Technology. Mt. Ranier, Maryland: Volunteers in Technical Assistance, 1979.

B. Subject Specific Thesauri

1. Agriculture and Nutrition

Baulkwill, W. Abridged Thesaurus of Land Resources Terms. Tolworth, Surrey, United Kingdom: Land Resources Division of the Directorate of Overseas Surveys, 1971, 3v.

FSTA Thesaurus. Shinfield, Reading England: International Food Information Service. n.d.

Manual of Classification of Agricultural and Forestry Research. Springfield, Virginia: National Technical Information Service, 1978. (Document No. PB 286099)

Thesaurus of Agricultural Terms. 2nd ed. Phoenix, Arizona: Oryx Press, 1977.

2. Education

Educational Resources Information Center. Thesaurus of ERIC Descriptors. 7th edition. Washington, DC: U.S. Government Printing Office, 1977.

Greaves, M.A. A Classification Scheme for Adult Education. London: National Institute of Adult Education, 1963.

UNESCO. IBE Education Thesaurus. Paris: UNESCO Press, 1973.

Viet, Jean. EUDISED: Multilingual Thesaurus for Information Processing in the Field of Education. 1st English ed. Paris: Mouton, 1973.

3. Energy

API Thesaurus. 15th edition. New York: American Petroleum Institute, 1978.

Cerny, B. Construction of a Suggestion List of Social Science Terms for the EDB Thesaurus: Therapy and Practice. Berkeley California: Information Methodology Research Project, Lawrence. Berkeley Laboratory, University of California, 1979.

Energy Information Data Base: Subject Thesaurus. Springfield, Virginia: National Technical Information Service, 1976. (Document No. TID-7000-R3)

Energy Microthesaurus - A Hierarchical Listing of Indexing Terms Used by NTIS. Springfield, Virginia: National Technical Information Service, 1976. (Document No. PB-254-800)

4. Management and Conservation of Natural Resources

Center Interprofessionel d' Etudes de la Pollution Atmospherique. Thesaurus Pollution Atmospherique. Paris, 1973.

Shriner, C.R., et al, eds. Environmental/Chemical Thesaurus. Oak Ridge, Tennessee: Oak Ridge National Lab, Union Carbide Corp. 1978. (also available through NTIS, (Document No. PB 284 310/OSL)

Environmental Microthesaurus - A Hierarchical List of Indexing Terms Used by NTIS. Springfield, Virginia: National Technical Information Service, 1977. (Document No. PB 265261)

Halpin, Peter, ed. Standard Air Pollution Classification Network: A Thesaurus of Terms. Springfield, Virginia: National Technical Information Service, 1979. (Document No. PB 292 038)

International Glossary of Hydrology. 1st ed. Geneva: World Meteorological Organization, 1974.

Microthesaurus of Air Pollution Terms. Philadelphia, Pennsylvania: The Franklin Institute of Research Laboratories, 1976.

Microthesaurus of Soil Mechanics Terms. Springfield, Virginia: National Technical Information Service, 1979. (Document No. AD-A003812)

Thesaurus of Water Resources Terms: A Collection of Water Resources and Related Terms for Use in Indexing Technical Information. 1st ed. Denver, Colorado: U.S. Bureau of Reclamation, 1971.

U.S. Office of Water Resources Research. Water Resources Thesaurus: A Vocabulary for Indexing and Retrieving the Literature of Water Resources Research Development. Washington, DC: U.S. Government Printing Office, 1966.

WES Microthesaurus of Scientific and Technical Terms. Springfield, Virginia: National Technical Information Service, 1977. (Document No. AD-A042 743/SL)

5. Manufacturing, Industry, and Services

Backer, S. Thesaurus of Textile Terms Covering Fibrous Materials and Processes. 2nd ed. Cambridge, Massachusetts. Massachusetts Institute of Technology, 1969.

Engineers Joint Council. Thesaurus of Engineering and Scientific Terms. New York, 1969.

Fasman, J. and Nobbs, P. Thesaurus of Pulp and Paper Terms. Quebec: Pulp and Paper Research Institute of Canada, 1965.

Textile Technology Terms and Information Retrieval Thesaurus. Charlottesville, Virginia: Textile Information Center, 1966.

United Nations. Department of Economic and Social Affairs. ISIC - International Standard Industrial Classification of all Economic Activities - New York: United Nations, 1968.

6. Medicine

Health Care Microthesaurus - A Hierarchical List of Indexing Terms Used by NTIS. Springfield, Virginia: National Technical Information Service, 1979. (Document No. PB-290776)

National Library of Medicine. Medical Subject Headings 1978. Bethesda, Maryland: National Library of Medicine, 1978.

7. Population Planning

Kolbe, H.K., and Bergman, R.F. Population/Fertility Control Thesaurus. Washington: Population Information Program, George Washington University, 1976.

Lucas, C. and Osburn, M. Population/Family Planning Thesaurus. Chapel Hill, North Carolina: Carolina Population Center, 1975.

Speert, K. and Wishik, S. Fertility Modification Thesaurus with Focus on Evaluation of Family Planning Programs. New York: Division of Social and Administrative Services, International Institute for the Study of Human Reproduction, Columbia University, 1973.

8. Transportation

Conference Europeenne des Ministres des Transporto Conference Centre de Documentatio.: Liste des Descriptors d'Economie des Transports. Paris: 1973, 3v.

9. Urban Planning

Boldt, R. and Charipar, G. Thesaurus Used in the Program in Urban and Regional Planning. Monticello, Illinois: Council of Planning Librarians, 1971.

Table 1

External Thesauri Ranked According to Criteria
Specified in User Survey

Requirement	Thesaurus						
	DIS	RANDD	UNESCO	SPINES	OECD	VITA	UNESCO STAT
Comprehensive Subject Coverage			X	X	X	X	X
Management and Technical Information Concepts				X			
Evaluation Terms				X			
Geographic Coverage				X	X		
Lead-In Vocabulary			X	X	X		
Standardization of Word Forms			X	X	X		
Specific Terms	X		X	X	X	X	
Logical Structure	X		X	X	X	X	X
Various Displays			X	X	X	X	
Translation			X	X	X		
Total of Matching Requirements	2	0	6	9	7	2	2

Exhibit B

Index Language _____

Outline for the analysis of thesaur/index languages/classification schemes

A Purpose (Ch. G, P, W)
Scope. Information system(s) in which to be used. Approach to indexing.
Type of file(s) and search mechanism(s) (printed index, on-line system, etc.) with which to be used.

B Nature, arrangement, and designation of descriptors (Ch. W, particularly W6)

B1 Degree of precombination and enumeration

B1a To what degree are descriptors precombined?

B1b To what extent are precombined descriptors enumerated and/or given in the alphabetical index? Built by the indexer? Updating characteristics.

B1c To what extent are core descriptors listed explicitly?

B2 How are descriptors designated?

B2a Core descriptors notation given /_/ only term given /_/

B2b Precombined descriptors: To what extent designated by an independent symbol, to what extent a string of symbols? Citation order free or fixed?

If independent symbol notation given /_/ only term given /_/

B3 How is the arrangement of descriptors determined?

B3a Arrangement of core descriptors. See A2a

B3b Arrangement of precombined descriptors

To what extent do the components of a precombined descriptor determine its place in the arrangement?

B4 Additional determinants for the sequence of descriptors.
(e.g., smallest to largest, steps of a process, time of evaluation)

C Content (concepts, terms, relationships)	D Form
	D0 Overall format Thesaurus parts and information given in each
C1 Plan of concepts	
C1.1 Coverage of concepts. Conceptual structure	
C1.2 Types and degree of differentiation of conceptual relationships. (Ch. Q). Completeness of conceptual relationship	D1.2 Display of conceptual relationships (Ch. Q and W) To what extent are the following devices used - arrangement - cross-references (give symbols used) - auxiliary ISAR system
C2 Plan of concept designations (Ch. O)	
C2.1 Coverage of terminology. Quality lead-in vocabulary	D2.1 Nature and quality of alphabetical index
C2.2 Types of terminological relationships. Completeness (are all synonyms and quasi-synonyms recognized?)	D2.2 Display of terminological relationships
C2.3 Nature of notations. (if none, state that)	

Total Matches Between Internal AID Sources and External Thesauri

Internal Sources	External Sources					
	UNESCO Thesaurus	Macrothesaurus	VITA Thesaurus	Draft Classification	Birch and Davis Proposed Codes	User Terminology
DIS Keyword List	369	287	127	55	18	110
RANDD Keyword List	1,282	863	246	139	20	159
Library of Congress Subject Headings	915	584	245	127	19	105
Activity Identification Purpose Codes	7	9	1	2	4	5
Activity Identification Technical Codes	50	44	20	28	19	23
Directory of Development Resources (Non-Thesaurus Terms)	110	72	47	11	8	20
Old Technical Codes	33	20	11	3	6	6
Objectively Verifiable Indicators	10	7	1	3	0	5
AID Document Distribution System (ADDS) Profile	68	54	19	25	17	19
Data Element Dictionary from the Economic and Social Data Bank	56	52	22	12	9	24
Business Classification Codes	5	5	7	18	4	5
Financial Classification Codes	35	25	18	7	6	8
Total	2,837	1,943	727	398	107	462

Table 3

Matched Terms as Percentage of
Total Terms in Each Source

External Source	Total Matched Terms	Total Terms in Source	Percentage of Matched Terms
UNESCO Thesaurus	2,837	10,269	27.6%
Macrothesaurus	1,943	3,799	51.1%
VITA Theasurus	727	1,602	45.3%
Draft Classification	398	4,661	8.5%

Table 4*

Matches of
DIS Keyword List with External Sources

External Source	Number of Matched Terms	Terms in AID Source	Percent
UNESCO Thesaurus	369	2,910	12.6
Macrothesaurus	287	2,910	9.8
VITA Thesaurus	127	2,910	4.3
Draft Classification	55	2,910	1.8

* Percentages in tables 4-12 do not add up to 100% as two other sources (Birch and Davis Proposed Codes and User Terminology) were also a part of raw frequency data. Both are irrelevant to this set of comparisons.

Table 5

Matches of
RANDD Keyword List with External Sources

External Source	Number of Matched Terms	Terms in AID Source	Percent
UNESCO Thesaurus	1,282	2,697	47.5
Macrothesaurus	863	2,697	31.9
VITA Thesaurus	246	2,697	9.1
Draft Classification	139	2,697	5.1

Table 6

Matches of
Library of Congress Subject Headings
with External Sources

External Source	Number of Matched Terms	Terms in AID Source	Percent
UNESCO Thesaurus	915	2,700	33.8
Macrothesaurus	584	2,700	21.6
VITA Thesaurus	245	2,700	9.0
Draft Classification	127	2,700	4.7

Table 7

Matches of
Activity Identification Purpose Codes
with External Sources

External Source	Number of Matched Terms	Terms in AID Source	Percent
UNESCO Thesaurus	7	404	1.7
Macrothesaurus	9	404	2.2
VITA Thesaurus	1	404	.2
Draft Classification	2	404	.4

Table 8

Matches of
Activity Identification Technical
Codes with External Sources

External Source	Number of Matched Terms	Terms in AID Source	Percent
UNESCO Thesaurus	50	331	15.1
Macrothesaurus	44	331	13.2
VITA Thesaurus	20	331	6
Draft Classification	28	331	8.4

Table 9

Matches of
Directory of Development Resources
(Non-thesaurus Terms) with External Sources

External Source	Number of Matched Terms	Terms in AID Source	Percent
UNESCO Thesaurus	110	338	32.5
Macrothesaurus	72	338	21.3
VITA Thesaurus	47	338	13.9
Draft Classification	11	338	3.2

Table 10

Matches of
Old Technical Codes with
External Sources

External Source	Number of Matched Terms	Terms in AID Source	Percent
UNESCO Thesaurus	33	132	25
Macrothesaurus	20	132	15.1
VITA Thesaurus	11	132	8.3
Draft Classification	3	132	2.2

Table 11

Matches of
Objectively Verifiable Indicators
with External Sources

External Source	Number of Matched Terms	Terms in AID Source	Percent
UNESCO Thesaurus	10	64	15.6
Macrothesaurus	7	64	10.9
VITA Thesaurus	1	64	1.5
Draft Classification	3	64	4.6

Table 12

Matches of
AID Document Distribution System
(ADDS) Profile with External Sources

External Source	Number of Matched Terms	Terms in AID Source	Percent
UNESCO Thesaurus	68	122	55.7
Macrothesaurus	54	122	44.2
VITA Thesaurus	19	122	15.5
Draft Classification	25	122	20.4

Table 13

Matches of
Data Element Dictionary
from the Economic and Social
Data Bank with External Sources

External Source	Number of Matched Terms	Terms in AID Source	Percent
UNESCO Thesaurus	56	309	18.1
Macrothesaurus	52	309	16.8
VITA Thesaurus	22	309	7.1
Draft Classification	12	309	3.8

Table 14

Matches of
Business Classification Codes*
with External Sources

External Source	Number of Matched Terms	Terms in AID Source	Percent
UNESCO Thesaurus	5	54	9.2
Macrothesaurus	5	54	9.2
VITA Thesaurus	7	54	12.9
Draft Classification	18	54	33.3

*Informal code list sent for inclusion in data base, July 1980.

Table 15

Matches of
Financial Classification Codes*
with External Sources

External Source	Number of Matched Terms	Terms in AID Source	Percent
UNESCO Thesaurus	35	171	20.4
Macrothesaurus	25	171	14.6
VITA Thesaurus	18	171	10.5
Draft Classification	8	171	4.1

*Informal code list sent for inclusion in data base, July 1980.