NETWORK RESEARCH BY DATA GRAPH MANAGEMENT FOR CAPACITY DEVELOPMENT AND KNOWLEDGE BUILDING IN SUSTAINABLE SANITATION

First D. Franquesa*, Second C. Amengual2

* Càtedra UNESCO de Sostenibilitat of the Universitat Politècnica de Catalunya
  EUETIT, Colom 1, 08222 Terrassa, Spain
  Email: david.franquesa@catunesco.upc.edu - Web page: http://www.catunesco.upc.edu

2 Àrea d’Ecologia Interdisciplinar of the Universitat de les Illes Balears
  Edif. Guillem Colom, 07122 Palma de Mallorca, Spain
  E-mail : caterina.amengual@uib.cat - Web page: http://uib.es

Key words: Sustainable Sanitation, Graph Databases, Capacity Development, Social Networks, Information storage and Retrieval

Summary. The Millennium Development Goals (MDG) provides clear targets by 2015 to, between others, ensure environmental sustainability. Water supply and sanitation were included in 2002 at the U.N. World Summit on Sustainable Development (WSSD) and it turns out that sanitation is by far the largest of all the MDG targets affecting about 40% of the global population. To be sustainable, a sanitation system has to be economically viable, socially acceptable and technically and institutionally appropriate and one objective of the Sustainable Sanitation Alliance (SuSanA) it’s to show how sustainable sanitation projects should be planned with participation of stakeholders with capacity development activities. Developing the capacity of individuals, organizations and societies to collaboratively learn through change and uncertainty is fundamental to environment and resource management and sustainability science. In knowledge management the social network analyses (SNA) it’s applied to help at the organizations to exploit their knowledge and the capacities distributed across their members. The aim of this contribution it’s to analyze the role of graph database management for improve capacity development and knowledge building in the sustainable sanitation frame through the individual, organizational, and environmental level. We provide a theoretical model with four features of network research: link analysis, social network, pattern recognition and keyword search. That affords to observe how the information in Sustainable Sanitation is scattered properly through the structure and also to detect the emergencies, objections and other characteristics of the network.
1 INTRODUCTION

In generic terms, sustainable systems are those which can be adapted to changing circumstances\(^1\). In order to be sustainable, a sanitation system has to be not only economically viable, socially acceptable, and technically and institutionally appropriate, it should also protect the environment and the natural resources\(^2\). Information and educational programs, introduction of new policies and regulations and capacity building and training of professionals are needed for sustainable sanitation.\(^3\)

The Millennium Development Goals (MDG) is an approach that provide targets promoted by the United Nations in 2000 to reduce poverty, hunger, illiteracy and others to ensure environmental sustainability. It was not until 2002 at the UN World Summit on Sustainable Development where water and sanitation where included as MDG7, target 10. Recently a pathway is formulated to achieve the target on sanitation through the sustainability. There are clear signs in the linkage between sanitation and all the other targets\(^4\).

Under this framework is necessary introduce the concept of ecological sanitation systems as safely recycle excreta and other organic waste products to crop production in such a way that the use of non-renewable resources is minimized. A change in the paradigm is necessary in the water management aspects. We need to end with the up to down vision, and start a learning process between different actors\(^5\). Social learning is needed to start a change towards adaptive management systems to sustainability\(^6\). Capacity development within sustainable sanitation it is a dynamic process where learning links up with live experience to improve outputs, processes and products\(^7\). Nowadays there is a lack on the outcomes monitoring when learning activities are done by different institutions at different levels. This field is taking up by the improvisation, but not always is working efficiently. How can engaging institutions record their outcomes in the learning approaches? There is some way to state the results achieved by knowledge provision? These are some of the question that we try to answer.

In the field of sustainable sanitation great efforts are leaded to capacity development and training with special emphasis on school sanitation\(^8\). The main objective is to work through different nodes and institutions all around the developing countries to test simple technologies that allow to bring sanitation to the poorest areas. The institutions in the north are giving appropriate knowledge and training to the nodes institutions in the south with the aim to help them to be advance in more sustainable practices. This process is suited in the learning-by-doing approach, where both partners are learning at the same time. How can their record the effect of the appropriate knowledge and the level on the competences in an area so intangible as social learning? To provide the world with more sustainable sanitation solutions will require enormous efforts in the area of capacity development. New tools are needed to evaluate the achievements made in this field.

This paper argues that if the information on Sustainable Sanitation resides on relational data model, this imposes difficulties for decision making based on exploration of the relationships among the data, such as paths, neighborhoods, patterns, and in definitive all queries based on entities that are interconnected satisfying a given constraint. That means a bad performance on time and cost. For improve this, we present a technology based on Graph
Database Models (GDMs) that implements efficiently four basic features of network research: link analyses, pattern recognition, social network and keyword search, as well as many applied research examples for decision making focused on capacity development and knowledge management activities.

2 BACKGROUND

2.1 Systems thinking

The holistic view, proposed among others by the principles of Bellagio, is addressed in systems thinking. Systems thinking is the attitude of the human being that is based on the perception of the real world in terms of totalities for their analysis, unlike the approach of the classical scientific method, which only sees the parts of this and so disjointed. Formally appears about 45 years ago, from the questions in the field of biology, Ludwing Von Bertalanffy questioned the application of scientific method to problems of biology, because this was based on a mechanistic and causal that was weak as the scheme for the explanation of the major problems that occur in living beings.

This question raised an intellectual paradigm to better understand the world around us, the emerging the paradigm of systems theory. Systems theory integrates in the analysis of situations and in the conclusions derived from them, suggesting solutions that are considered to have several elements and relationships that form the structure of which is defined as a "system" and everything that makes up the system environment defined. The basic philosophy behind this position is the holism from the Greek \( \text{holos} = \text{whole} \).

Better understand the interlinkages between social and natural systems will help us take the appropriate action to act in coherence with the natural system, this is a concept linked to resilience\(^{10}\). Sustainable Development would be one in which disturbances generated from the social system could be 'embedded' with the natural without assuming a change of state which, in turn, affect strongly the dynamics of the social system. Achieving this goal requires action types: Coping with the impact of the shocks produced by the social system (trying to 'sync' human activities with the cycles of natural subsystem), and increase the capacity of the two subsystems to adapt to shocks.\(^{11}\) Learning provides a basis for the joint action required to respond to social–ecological feedback.\(^{12}\)

In the middle of last century, authors as Durkheim argued that human societies were like biological systems in that they were made up of interrelated components. As such, the reasons for social regularities were to be found not in the intentions of individuals but in the structure of the social environment in which they are embedded\(^{13}\). How decision support systems will help us on making decisions if the underlying data model is too much structured and does not act effectively when our data are not structured. It seems that the theory of Ashby\(^{14}\) is still in effect, he proposed his Law of Requisite Variety, which stated that a model (i.e. a representation) can only represent some aspect of reality if it has sufficient internal variety to capture the complexity of that reality.
2.2 Network analysis for Sustainable Development

In knowledge management social network research has been successfully with the aim of to help organizations better exploit the knowledge and capabilities distributed across its members. A network map shows the nodes and links in the network. Nodes can be people, groups or organizations. Links can show relationships, flows, or transactions. A network map is an excellent tool for visually tracking your ties and designing strategies to create new connections, and also excellent ‘talking documents’ – visual representations that support conversations about possibilities.

Network research is hot today, with the number of articles in the Web of science on the topic of "social networks" nearly tripling in the past decades. It affords to explain social phenomena in different approaches. Whereas traditional social research explained and individual's outcomes or characteristics as a function of other characteristics of the same individual (e.g., income as a function of education and gender), social network researchers look to the individual's social environment for explanations, whether through influence processes (e.g., individuals adopting their friends' occupation choices) or leveraging processes (e.g., an individual can get certain things done because of the connection he has to powerful others).

For analyzing intensity of relations among stakeholders social network analysis can be used. Networks provide a broad and inclusive framework: i) Networks can be described and analyzed at many scales, from interactions between individuals in small rural communities to international linkages between large organizations, ii) There is a range of tools available to describe and measure networks, which is relevant to the analysis, planning and evaluation of change in those networks, iii) There is an extensive and developing body of theory and research on the nature of networks, that spans many disciplines, and which is available to help inform development agencies’ theories of change.

There are other areas of research and theorizing about social networks that have relevance to development aid projects. In the health sector there is already an established record of social network analysis techniques being used as part of epidemiological studies, as well as in studies of the effectiveness of health communications, especially in the field of HIV/AIDS.

2.3 Why a graph data model: Limitations of relational data model for the study of interlinkages

The term “data model” or "database model" has been widely used in the information management community: it covers various meanings. In the most general sense, a database model is a collection of conceptual tools used to model representations of real-world entities and the relationships among them. The term is also often used to refer to a collection of data structure types.

The differences between graph data models and the relational data model are manifold. For example, the relational model is geared towards simple record-type data, where the data structure is known in advance (airline reservations, accounting, inventories, etc.). The schema is fixed, which makes it difficult to extend these databases. It is not easy to integrate different
schemas, nor is it automatable. The query language cannot explore the underlying graph of relationships among the data, such as paths, neighborhoods, patterns.\(^\text{19}\) The relational data model is now more than 30 years old. It's good for a number of scenarios and can handle certain types of data very well. But it isn't perfect. For data that is semistructured and/or network oriented, the relational database offers poor runtime characteristics. Furthermore, it forces a static development cycle and is of little help to those who have to live with a domain model that is constantly changing, even after deployment. This translates to wasted development time and money.

Generalizations of the classical database models\(^\text{20}\); classical model were criticized for their lack of semantics, the flatness of the permitted data structures, the difficulties the user has to “see” the data connectivity, and how difficult it is to model complex objects\(^\text{21}\).

Graph database models are applied in areas where information about data interconnectivity or topology is more important, or as important, as the data itself. In these applications, the data and relations among the data, are usually at the same level. Introducing graphs as a modeling tool has several advantages for this type of data:\(^\text{19}\)

— It allows for a more natural modeling of data. Graph structures are visible to the user and they allow a natural way of handling applications data, for example, hypertext or geographic data. Graphs have the advantage of being able to keep all the information about an entity in a single node and showing related information by arcs connected to it\(^\text{22}\). Graph objects (like paths and neighborhoods) may have first order citizenship; a user can define some part of the database explicitly as a graph structure\(^\text{23}\), allowing encapsulation and context definition\(^\text{21}\).

— Queries can refer directly to this graph structure. Associated with graphs are specific graph operations in the query language algebra, such as finding shortest paths, determining certain subgraphs, and so forth. Explicit graphs and graph operations allow users to express a query at a high level of abstraction. To some extent, this is the opposite of graph manipulation in deductive databases, where often, fairly complex rules need to be written\(^\text{23}\). It is not important to require full knowledge of the structure to express meaningful queries.\(^\text{24}\) Finally, for purposes of browsing it may be convenient to forget the schema.\(^\text{25}\)

— For implementation, graph databases may provide special graph storage structures, and efficient graph algorithms for realizing specific operations.\(^\text{23}\)

3 FRAMEWORK: DEX TECHNOLOGY BASED ON GDM

In order to find that our theoretical model in practice work, we chose a technology that assures a good behavior and performance with our problem.

In this section we present DEX as a possible technology based on GDMs, developed for DAMA-UPC\(^\text{26}\). DEX is basically characterized by three properties: data structures are graphs or any other structure similar to a graph; data manipulation and queries are based on graph-oriented operations; and there are data constraints to guarantee the integrity of the data and its relationships.
3.1 What does Graph Database Models contribute to the problem solution

The development of huge networks such Internet, geographical systems, transportation or automatically generated social network databases, has brought the need to manage information with inherent graph-like nature. In these scenarios, users not only keen on retrieving plain tabular data from entities, but also relationships with other entities using explicit or implicit values and links to obtain more elaborated information. In addition, users are typically not interested in obtaining a list of results, but a set of entities that are interconnected satisfying a given constraint.

Cases like bibliographic database are a clear example where a more complex querying system would be beneficial. In these scenarios, the user might not be only interested in finding an specific author or publication, but to analyze the relationships within a group of authors or publication, to understand the relevance of an specific paper or any other implying the exploration of the relationships between entities.

Those environments impose three important problems: (i) the continuous growth of the data sources, (ii) the need for a versatile querying system that allows Information Retrieval (IR) queries with different flavors ranging from keyword search to the complex mining of patterns in graphs, and (iii) the need to integrate data coming from different sources to enrich the answers to complex queries over incomplete databases.

Data of EcoSanRes is highly-interconnected and has a complex structure, which is difficult to capture in any current database (Excel files, Forums, many databases, etc.). EcoSanRes information system is based on relational model, this model are well suited for queries based on values, like equalities or range search, but in these models the exploration of relationships must be always set explicitly by joins even if foreign keys have been declared, and it becomes really difficult to explode all the potential relationships of a node. Not only this: for self-relationships, the relational queries require recursive extensions that are more difficult to create and manage. On the opposite side, the natural mechanism of GDMs is the automatic exploration of the relationships in a graph, represented in the form of edges between nodes. It is a relationship rather than a value oriented analysis.

However, in a real scenario with complex analysis where the source data does not represent the semantics of the query (e.g., multiple data integrated for keyword search), or where some hidden patterns must be detected in the structures (e.g., social network of a node), more sophisticated query process is required.

In conclusion it seems that the problem of EcoSanRes and many other organizations that fall into their database, it isn’t designed to integrate data from multiple sources, or to respond efficiently queries based on entities that are interconnected satisfying a give constraint, and other issues that are not come by default in relational model. This implies that the questions that need answers or you sense that can bring knowledge, are significant investments of time and money, it means change the entire system, from the initial schema of the database to user queries.
3.2 Data representation

DEX deals with two different types of graphs: the DbGraph, a single graph that contains the schema and data extracted from external data sources and stores them in a persistent storage; and the RGraph, the result of DEX query, which is stored in the temporal storage. The information contained in the DbGraph is used as source data by DEX queries to obtain the results in the form of one or more RGraphs. The DbGraph includes two subgraphs: the schema subgraph, which contains the metadata of the data sources, and the data subgraph, which contains the data loaded from the data sources as defined in the schema subgraph. Both are connected through edges between the entity definitions of the schema subgraph and their instance objects in the data subgraph. The schema subgraph contains the following node types:

- **datasource**: contains information of the data source and its type (ODBC, CSV, XML, WEB, etc.). It is connected to one or more datasets.
- **dataset**: the definition of an entity or collection of data units (rows, XML elements, HTML pages, etc.). It is defined by one or more attributes.
- **attribute**: a characteristic or property of a dataset (column, xml attribute, etc.). It has a name and a data type and can be part of one or more relationships.
- **relationship**: it can represent either any edge between two instances in any dataset, or the definition of a constraint between two attributes (for example a foreign key).

Our theoretical data model contains a set of interrelationships that are necessary. For example to give answers about: what is the social network of people directly and indirectly linked to a particular project, the research being done on Sustainable Sanitation and whether it will cover all nodes of the organization, what relationships have been established after training in ecological sanitation.

Figure 1: E/R representation shows an example of the mapping between a simple RDBMS, and Figure 2: DbGraph for a Ecosanres the corresponding schema subgraph through it DEX representation. Attributes belong to dataset and datasets to datasources, creating a hierarchy that depends on the original data sources. However, relationships may link datasets from a single or across multiple data sources (intra- or interrelationships, respectively). This goes beyond the classical data models, allowing the DbGraph to connect and relate heterogeneous data sources that share common identifiers, as if they were foreign keys across different data sources.

![Figure 1: E/R representation](image)

Figure 1: E/R representation
Figure 2: DbGraph for a Ecosanres

Keys are Sustainable Sanitation activities. Persons are individuals who have participated doing an specific role (technicals, responsibles, authors, etc.) in such activity. Thus, if there are two distinct databases that have the entity PERSON, and both entities have a public person identifier like email as a primary key, a relationship between them can be created within the DbGraph to join all the information of each single person. This provides the illusion of a single and integrated storage.

DbGraph allows for the integration of multiple data sources mapped from different data models. This data sources can be relational database, tabular files with rows and columns, XML documents, HTML pages, RDF graphs, etc. For each different data model there exists and specific mapping to the Database.

4 THEORETICAL MODEL WITH DEX TECHNOLOGY

The network offers numerous resources and communication channels for EcoSanRes learning nodes. The resource format is diverse, so are videos that cover a specific problem, such as forums where experts to advice staff that is carrying out projects in the nodes. The data from our system come from different bakeries (e.g. db, Forums, Emails, ...). We have to collect all this data and give to the users the possibility to analyze it. The main drawback of EcoSanRes is interlace all this information, place it in the nodes and provides an impact assessment that causes this information when reach on-line to the users or on-site trainings.

4.1 EcoSanRes data Model proposal

The following Figure 3 contains projects, forums, films, articles and other information related to the Sustainable Sanitation (SS). The most important data sets are PERSON, which contains all individuals involved in any key, KEY, which contains basic information of each activity on capacity building or knowledge management, ROLE_P, which contains who (PERSONS) and how (ROLE_TYPE) are they participated into a key, LINK, which contains the relations between keys using different types of links.
Figure 3: E/R Diagram of Sustainable sanitation

Under the name of the entity appear the number of objects that holds the entity, it has been taken from similar E/R Diagram where the query's performance has been tested, e.g. about two million people involved in sanitation issues.

This is a list of instances of entities of the figure. KEY_TYPE holds a set of tools for capacity building that can be found in sustainable sanitation.

<table>
<thead>
<tr>
<th>KEY_TYPE</th>
<th>ROLE_TYPE</th>
<th>KEY</th>
<th>LINK_TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video</td>
<td>Author paper</td>
<td>Info_id</td>
<td>Cites</td>
</tr>
<tr>
<td>Case study</td>
<td>Project Trainer</td>
<td>Title</td>
<td>Answer</td>
</tr>
<tr>
<td>Project</td>
<td>Project Responsible</td>
<td>Text</td>
<td>Equivalente</td>
</tr>
<tr>
<td>Training</td>
<td>Project manager</td>
<td>Url</td>
<td>Recommend</td>
</tr>
<tr>
<td>movie</td>
<td>Technical Expert</td>
<td>Date</td>
<td>Responds to</td>
</tr>
<tr>
<td>Paper</td>
<td>Map</td>
<td></td>
<td>Arises from</td>
</tr>
</tbody>
</table>

Figure 4: Instances of E/R Diagram

4.2 Example of application on capacity development and knowledge building

We explain an example that uses four kind of features: link analysis, showing that is able to explore the relationships between objects, social networks, to analyze the relationships between groups of affine objects, pattern recognition, showing the ability to identify patterns into the relationship links between entities, and, finally, keyword search, showing how to find relationships between objects that share, in some way, a given keyword, and returns the context of such objects.

We explain a case where our proposal of information system can result appropriate for this objective. Imagine that we are responsible for Sustainable Sanitation (SS) node project coordination; we have to develop a project with a very tight budget, we live in a city of Africa
quite distant with experts on SS. In addition, our staff doesn’t have all the necessary skills to develop the project.

We need to find an expert who can do training to our staff in the need it skills. We use pattern recognition feature (identify patterns into the relationship links between entities). If our database has relations between skills and personnel, this can be done easily in classical relational databases, but we don’t want this kind of relations, we want to manage facts on SS (KEYS); somebody has done a project, has written about something, has been recommended for an expert, etc. The feature explores each of the KEYS done by each expert, identifying if the KEYS done by the experts are indirectly linked with skills.

The result would be a set experts names. Before contacting them, we decided to investigate them to validate the quality of related KEYS they have done. In this case we use link analysis feature, such as: Give me all the information about John Smith in these skills. If we want we can to specify more restrictions (country, year, etc.) for narrowing the search

After validating skills: projects, work, published books, etc., we want their referrals. Who does we ask? Then we make a Social Network query that will return people who have been in relationship with someone of Sustainable sanitation network. Now we could ask for references or ask our colleagues to contact them.

To complete the case, we assume that we do not find any domain expert in the skills asked. At this point we have to find an expert, the question is where to look. Again we want to facilitate the work with a new query. We ask the system to information sources that address issues related to the skills we need, we specify some details of the search: just communication channels of information like mailing lists, forums or websites. The result is a graph where nodes are skills and communication channels (forums, mailing lists, contacts, etc ...) and edges represent the which skills deals with each communication channels. Once we have the emails we send them the offer of employment. Note how the query in the reverse also may be interesting: from a set of distribution lists to a set of skills, tags, etc. adressed.

### 4.3 Requirement Analysis

The objective of this section is to verify the capability of the model to solve different kinds of queries, not only from the performance point of view, but also trying to answer real complex questions in a database.

Some examples grouped by four features, essential to allow members of Sustainable Sanitation to study the impact of its activities in capacity development, four features are: a) Link analysis, b) Social networks, c ) Pattern recognition, d) Keyword search.

For each query we present a result in a “map network” form and the processing time required, time is an approximation, since the information presented is a theoretical model, but the queries has been tested by DEX in a similar E/R data model, that we call IMDb (Internet Movie Database).^{27}

#### 4.3.1 Link Analysis

We are interested on exploring the relations between the nodes of the graph, navigating the edges between them.
Query 1 (Q1): Get all the information of a movie m (KEY_TYPE = movie)

An example of link analysis is Q1, where all the information of a movie is obtained in a single graph. The root node will be a node from the data set KEY with KEY_TYPE = "movie". From this initial node, we explode a graph containing all the ROLE_P and ROLE_O (Role of persons and organizations in Sanitation movies), persons, information of the tags used, including the chain of information referenced to the root by LINK, that can be any kind of KEY_TYPE. Note that this could pose a serious problems to traditional relational systems that would have to resort to recursive queries including a large number of join operations, increasing significantly the complexity of such queries. 27. A simplificated example of information retrieved for movie “Water and sanitation in NIGER challenges” is:

```
Figure 5: movie link analysis
```

In IMDb, for a query related to movie 'Crash'. The resulting graph for in this movie has 2054 nodes and 3692 edges, and it takes less than 0134 seconds to be obtained. 27

4.3.2 Social Networks

Previous query show that DEX can be used to explode the links between different entities in a graph. Now, we depict an example where DEX is used to analyze social networks. In this case the focus on the relationship between different groups of nodes with the same affinity.

Let us to consider all the persons who have participated in the same project to form a group in a social network. Specifically in this case, we define partnership as the relationship between two persons who have done it the same project. We impose two restrictions. First, we restrict the explored KEYS to only those KEYS corresponding to a project (KEY_TYPE = "Project"). Second, we restrict the person role to "technical". We apply this conditions because (i) we want to increase the query complexity rather than always exploring everything and, (ii) our database contains a lot of information (forums, article,etc.) that could provide unrealistic relationships between persons.

Query 2 (Q2): Find the minimum collaboration distance between two technical experts.
The first query exploring partnerships, Q2, tries to find the minimum distance between two persons that has worked as a technical. If distance is 1 it means that both have worked in the same project; a distance of 2 means that they never worked together, but exists at least another partner who has done a project with both of them.

Q2, in IMDb is executed in less than 1.52 seconds when the memory is not limited and involves near 110000 operations.

**Query 3 (Q3): Find the full relationships network of all the partners of a technical expert.**

Q3 is a more complex query. Instead of looking for relationships between two experts, we are now interested in knowing the full network of relationships of all the partners of one expert. This kind of analysis provides us with a lot of information about partnership patterns, groups of technician based on the sanitation agencies. In IMDb, after 6 minutes of execution, DEX returned a dense graph with 1051 partner’s nodes and near 5.5 million partnerships edges.

### 4.3.3 Pattern recognition

Pattern recognition defines a different kind of queries, where a lot of potential graphs can be created and explored, but only a few of them will qualify because they mach a certain pattern.

**Query 4 (Q4): Find all the responsible that have worked with the same technical expert in three different projects made in a period of time of five years**

As an example of pattern recognition, Q4 tries to find all the coordinators with three projects in less than five years with the same technical expert, i.e., it would be some kind of 'muse' detector query. This is a complex query that not only requires pattern detection but also involves several data filters like coordinator role, technical expert role or key_type equal to project. In IMDb context, a similar query required more than 6 million operations in near 6.5 minutes. See Figure 6

![Figure 6: ‘muse’ detector](image-url)
4.3.4 Keyword search

We have shown the ability of DEX to return a set of results when the user knows the schema or at least part of it, i.e., we assume that the user has knowledge on how the data is structured. However, this assumption may be unrealistic in some scenarios like the WEB or documental database. In this last query we show that DEX is also suitable to perform a keyword search, where the user is assumed not to know anything about the organization of the data.

Note that the most of keyword search engines are based on data storage where data has been tokenized and indexed to provide fast retrieval of documents related to keywords. In conventional databases models like the relational model, this time of search requires a full indexing of all the string columns and it becomes unfeasible due to the high cost in terms of storage size and performance. DEX can take advantage of dictionaries and compressed structures\(^\text{27}\). Obviously, the changes for DEX to perform better than a documental or a tokenized storage are low. In addition, DEX does not only return a list of entities including the set of keywords but an entity-context graph containing interesting extra information for that entity, enriching the quality of the results.

**Query 5 (Q5): Return all the context information of all the entities containing the tag Biogas production, expert John Smith and County India**

Such queries are done separately in different information sources, db queries, queries in forums, consultations db project. EcoSanRes doesn't has a tool that integrates all sources in one query, in spite of the use of Internet search engines, but narrow the search to the semantic domain of sustainable sanitation is difficult or isn't possible, and the format of results presented to the user is only a list. The present potential for the user is the ability to perform a search across different data sources and show the results in a visual format that allows the user to navigate easily. The Figure 7 shows the result obtained with word "Biogas production". We run Q5 in IMBD context and it requires only 1.9 seconds.\(^\text{27}\)

\[\text{Figure 7 Biogas production search}\]
Query 7 (Q7): How many practitioners are doing Sustainable Sanitation after a training course?

This query gives the effectiveness of training. It is considered that a "practitioner" does SS if carried out SS activities (KEY = projects, etc..) after the SS training. We restrict to the TAGs introduced during the training because we want to know if they are applying this concepts, technologies, etc. These are the following steps to build the query.

1) It is consulted the persons who have participated in a SS training.
2) For all persons of step 1) are added to a network their KEYs and TAGs, with date after the training, the result would be a network of KEYs, TAGs, PERSONS, see Figure 8. Tag 3 and 4 are belongs to Sustainable Sanitation.
3) From this network, a new network is builded connecting persons and SS activities (KEY) done after the training, then are added the persons of the first step. People not connected have not participated in any of Sustainable Sanitation project but were in training. See Figure 9.

As an indicator of the training effectiveness we use inclusiveness operation “Inclusiveness of a graph is the total number of points minus the number of isolated points”. In our case the percentage of training effectiveness is 75%. You can export the resulting graph to other software, as UCINET, to obtain this type of indicators.

We could do a new version of the query to see the emergence of new knowledge and transmission, e.g. in the first case new TAGS that not belong to the training appear or in the second new people who have never been trained.

---

2 Network of people, key and tags. Tags 3 and 4 belongs to SS, so that the keys 1 and 2 are considered as SS activities, as a result, persons 1, 3 and 4 had participated in a SS training.
CONCLUSIONS

In this article we deal with problems in where network research allows adapting the training needs and access to appropriate knowledge on Sustainable Sanitation.

- We have exposed how to deal with complex questions, as: How many projects or activities are resulting from a training course, Who is an expert in this skills, What a community of practitioners are doing in Sustainable Sanitation, What kind of information is related to these tags, etc.

- Can be effective for someone who is just beginning on Sustainable Sanitation or other areas of sustainability, features as keyword search don’t require knowing the details or the schema of our database. This can help to give autonomy to the southern countries in the process of finding information and experts.

- Simplifies the process of searching for information on the Internet, it allows us to retrieve multiple external data sources (database, HTML pages, etc..), integrate with internal (CSV, XML, relations databases), and present the information.

- We have seen the advantages that offer the features of: link analyses, pattern recognition, social network and keyword search. These features can be efficiently used to manage and evaluate more easily the capacity development and knowledge building process.

REFERENCES


[26] DAMA-UPC, the DAta MANagement group at Universitat Politécnica de Catalunya (UPC) is part of the Computer Architecture Department (DAC). http://www.dama.upc.edu
