**Change History**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Version | Date | Reason | Author | Approved |
| 1.0 | 19th May 2010 | Initial Version | Jane Harrison  Xing-Shu Liu |  |
| 1.1 | 26th Aug 2010 | Minor changes of ERD | Chih-Hsiang Tang |  |
| 1.2 | 3st Dec 2010 | Update of database size estimate | Chih-Hsiang Tang |  |
| 1.3 | 27th Sep 2010 | Sampler Media now stored in files; DB size estimate will remain the same and assume that files are part of the database | Xing-Shu Liu |  |

Contents

[1. Entity Relationship Diagram 2](#_Toc263252887)

[2. TiGERS Data Mapping 3](#_Toc263252888)

[3. TiGERS Relational Schema 4](#_Toc263252889)

[4. Database Size Estimate 5](#_Toc263252890)

[4.1. Introduction 5](#_Toc263252891)

[4.2. Assumed Space Usage and Allocation 5](#_Toc263252892)

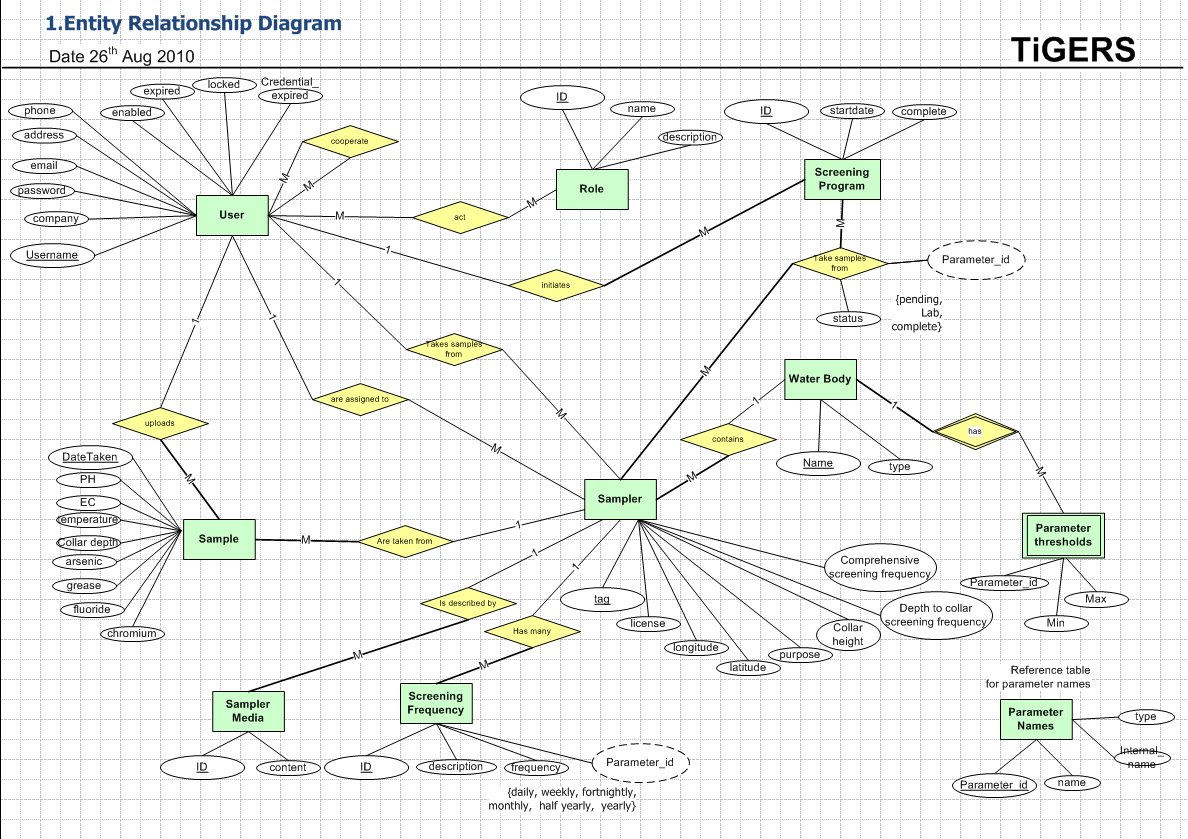
[4.3. Estimated Number of Records 6](#_Toc263252893)

[4.4. Estimated Database Size 7](#_Toc263252894)

[4.5. Database Size Estimate Worksheet 8](#_Toc263252895)

[4.6. Miscellaneous Notes 15](#_Toc263252896)

# Entity Relationship Diagram



# TiGERS Data Mapping

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| User: | username | company | password | phone | email | address | enabled | expired |
|  | locked | credential\_expired |  |  |  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Cooperate | id | contractor\_name | laboratory\_name |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ScreeningProgram: | id | start\_date | complete | username |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sample: | date\_taken | PH | EC | temperature | collar\_depth | arsenic | grease |
|  | fluoride | chromium | username | tag |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sampler: | tag | license | | longitude | latitude | collar\_height | | comp\_screening\_freq | purpose |
|  | contractor | | waterbody | | laboratory | | depth\_to\_collar\_screening\_freq | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SamplerMedia: | id | file | description | tag |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ScreeningFrequency: | id | description | frequency | tag |

|  |  |  |
| --- | --- | --- |
| FrequencyItem: | frequency\_id | parameter\_id |

|  |  |  |
| --- | --- | --- |
| Waterbody: | name | type |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ParameterThresholds: | name | parameter\_id | Min | Max |

|  |  |  |  |
| --- | --- | --- | --- |
| ParameterNames: | parameter\_id | name | Internal\_name |

|  |  |  |  |
| --- | --- | --- | --- |
| ScreeningItem: | program\_id | tag | parameter\_id |

|  |  |  |  |
| --- | --- | --- | --- |
| ScreeningProgamSamplers: | id | tag | status |

# TiGERS Relational Schema

|  |  |  |  |
| --- | --- | --- | --- |
| User:  username  company  password  phone  email  address  enabled  expired  locked  credential\_expired | varchar(20) not null  varchar(50) not null  varchar(30) not null  char(10) not null  varchar(60) not null  varchar(100) not null  char(1)  char(1)  char(1)  char(1) | ScreeningProgram:  id  start\_date  complete  username | integer autoincrement  date not null  char(1)  varchar(20) not null |
| Sample:  date\_taken  PH  EC  temperature  collar\_depth  arsenic  grease  fluoride  chromium  username  tag | date not null  number(5,2)  number(5)  number(5,1)  number(5)  number(5,3)  number(3)  number(5,3)  number(5,3)  varchar(20) not null  varchar(5) not null | Sampler:  tag  license  longitude  latitude  collar\_height  comp\_screening\_freq  purpose  contractor  waterbody  laboratory  depth\_to\_collar\_screening\_freq | varchar(5) not null  varchar(20)  number(9,6)  number(9,6)  number(5)  varchar(15) not null  varchar(50) not null  varchar(20)  varchar(30) not null  varchar(20)  varchar(15) |
| SamplerMedia:  id  file  description  tag | integer autoincrement  varchar(100) not null  varchar(200) not null  varchar(5) not null | FrequencyItem:  frequency\_id  parameter\_id | integer not null  integer not null |
| ScreeningItem:  program\_id  tag  parameter\_id | integer not null  varchar(5) not null  integer not null | ScreeningProgamSamplers:  id  tag  status | integer  varchar(5) not null  char(1) not null |
| ParameterThresholds  name  parameter\_id  Min  Max | varchar(30) not null  integer not null  number  number | ParameterNames:  parameter\_id  name  internal\_name  type | integer autoincrement  varchar(20) not null  varchar(10) not null  char(1) not null |
| ScreeningFrequency:  id  description  frequency  tag | integer autoincrement  varchar(20) not null  varchar(15) not null  varchar(5) not null | Cooperate:  id  contractor\_name  laboratory\_name | Integer autoincrement  varchar(20) not null  varchar(20) not null |
| Waterbody:  name  type | varchar(30)  char(1) not null |  |  |

# Database Size Estimate

## Introduction

The following will provide a rough estimate on the expected size of the database for all records once they have been converted from the old system, as well as estimates for continued usage into the future.

Estimations consider the following:

* Data type of each field and the disk space usage of each
* Estimated average length for variable length fields (such as text or VARCHAR2)
* Total size of each database table record (or row), including allocation for DBMS overhead
* Total number of records in each table, for current and future usage
* Total size of each table, based on the total size of each record and the number of records in the table, as well as allocation for DBMS overhead
* Allocation for indexes (other database objects such as triggers not considered here; the impact on size they have should be negligible anyway). For this estimate, only primary keys will be indexed.
* Additional “safety” allocation to increase probability that size is within estimated range

Note that “DBMS overhead” refers to allocation allowed for overhead which the database management system (DBMS), Oracle, may impose (for example, for metadata, or optimisation purposes)

## Assumed Space Usage and Allocation

The estimate will need to assume the size and overheads for various items. This estimation will use the following figures:

|  |  |
| --- | --- |
| Item | Space Usage |
| Number (decimal) field | 8 bytes per value |
| Boolean field | 1 byte (type set as char(1)) |
| Integer field | 4 bytes per value |
| Date field | 8 bytes per value |
| Char/text field | Length is equal to the defined field size, e.g. char(10) uses 10 bytes |
| Varchar field | Length is 1 byte per character (average length), plus 1 byte for length of text |
| Per-row overhead | Additional 25% of size of row |
| Per-table overhead | 32 bytes per field, plus additional 30% of size of table |
| Indexes | 50% of the data being indexed |
| Overall safety allocation | Additional 20% of total size |

## Estimated Number of Records

The estimation will also need to make an assumption on the number records which will exist in each table after initial data conversion, as well as potential growth into the future.

For this estimate, we will assume 20 years of historical data will be converted and stored in the database (“now”), and make provisions for the likely size in another 20 years’ time (“future”).

The following table shows the estimated number of records for “now” and “future” and expected update frequency:

|  |  |  |  |
| --- | --- | --- | --- |
| Table | Now | Future | Notes |
| Contractor | 10 | 20 |  |
| Laboratory | 20 | 40 | Seems to be an implication that number of laboratories exceeds number of contractors |
| Cooperate | 20 | 40 |  |
| EnvironmentalOfficer | 5 | 10 |  |
| User | 35 | 70 | Sum of contractors, laboratories and environmental officers |
| Waterbody | 40 | 80 | There are probably more water bodies than laboratories |
| Sampler | 200 | 400 | Assuming an average of 5 samplers per waterbody |
| SamplerMedia | 800 | 1,600 | Assuming an average of 4 images/videos per sampler (since there’s a maximum of 10) |
| Sample | 1,461,000 | 3,652,500 | Assuming that all the 200 samplers have existed for whole 20 years and records samples daily: 365.25 \* 20 \* 200 = 1461000. For future, assume 200 samplers for another 10 years and then 400 samplers for 10 years after that = 1461000 + 365.25\*10\*200 + 365.25\*10\*400 = 3652500 |
| ParameterThresholds | 320 | 640 | Assume each waterbody has a threshold for each parameter |
| ParameterNames | 8 | 8 | Currently only sampling 8 parameters. Changing this number will require changes to the Sample table. This estimation will assume this number remains constant. |
| ScreeningProgram | 100 | 400 |  |
| ScreeningProgramSamplers | 150 | 600 | Will contain the same or more records than ScreeningProgram table |
| ScreeningItem | 600 | 2400 | Assuming 4 parameters on average over ScreeningProgramSamplers |
| ScreeningFrequency | 100 | 400 | Assuming similar to ScreeningProgram |
| FrequencyItem | 400 | 1600 | Assuming 4 parameters on average over ScreeningFrequency |

## Estimated Database Size

With the above, size estimates for each table can be derived, and summed up to give the estimated database size.

Current space usage: 5,705.20 MB  
Future space usage: 11,533.03 MB

Estimations include a 20% safety margin. Full calculation details can be found in section 4.5.

## 

## Database Size Estimate Worksheet

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| User | | |  |  | ScreeningProgram | | |  |
| **Fields** |  | | Avg. field size (bytes) |  | **Fields** |  | | Avg. field size (bytes) |
|  | username | | 9 |  |  | id | | 4 |
|  | name | | 13 |  |  | start\_date | | 8 |
|  | password | | 21 |  |  | complete | | 1 |
|  | phone | | 10 |  |  | username | | 9 |
|  | email | | 26 |  |  | *Total* | | *22* |
|  | address | | 30 |  | **Projected avg. row size** | | | 27.5 |
|  | enabled | | 1 |  | **Avg. num rows** | |  | 10 |
|  | expired | | 1 |  | **Size of table (data)** | |  | *275* |
|  | locked | | 1 |  | **Fields overhead** | |  | *128* |
|  | credenial\_expired | | 1 |  | **Index overhead** | |  | *20* |
|  | *Total* | | *113* |  | **Projected table size** | |  | **550** |
| **Projected avg. row size** | | | 98.75 |  |  | |  |  |
| **Avg. num rows** | |  | 70 |  |  | |  |  |
| **Size of table (data)** | |  | 6912.5 |  |  | |  |  |
| **Fields overhead** | |  | 320 |  |  | |  |  |
| **Index overhead** | |  | 329 |  |  | |  |  |
| **Projected table size** | |  | **9073.8** |  |  | |  |  |
|  | |  |  |  |  | |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sample |  | |  |  | Sampler |  |  |
| **Fields** |  | | Avg. field size (bytes) |  | **Fields** |  | Avg. field size (bytes) |
|  | date\_taken | | 8 |  |  | tag | 6 |
|  | PH | | 8 |  |  | license | 11 |
|  | EC | | 8 |  |  | longitude | 8 |
|  | temperature | | 8 |  |  | latitude | 8 |
|  | collar\_depth | | 8 |  |  | collar\_height | 8 |
|  | arsenic | | 8 |  |  | comp\_screening\_freq | 11 |
|  | grease | | 8 |  |  | purpose | 16 |
|  | fluoride | | 8 |  |  | contractor | 9 |
|  | chromium | | 8 |  |  | waterbody | 11 |
|  | username | | 9 |  |  | laboratory | 9 |
|  | tag | | 6 |  |  | depth\_to\_collar\_screening\_freq | 6 |
|  | *Total* | | *87* |  |  | *Total* | *103* |
| **Projected avg. row size** | | | 108.75 |  | **Projected avg. row size** | | 128.75 |
| **Avg. num rows** | |  | 1461000 |  | **Avg. num rows** |  | 200 |
| **Size of table (data)** | |  | 158883750 |  | **Size of table (data)** |  | 25750 |
| **Fields overhead** | |  | 352 |  | **Fields overhead** |  | 352 |
| **Index overhead** | |  | 5844000 |  | **Index overhead** |  | 600 |
| **Projected table size** | |  | **214146533** |  | **Projected table size** |  | **34713** |
|  | |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ParameterNames | |  |  | FrequencyItem | |  | |  |
| **Fields** |  | Avg. field size (bytes) |  | **Fields** | |  | | Avg. field size (bytes) |
|  | parameter\_id | 4 |  |  | | frequency\_id | | 4 |
|  | name | 8 |  |  | | parameter\_id | | 4 |
|  | internal\_name | 4 |  |  | | *Total* | | *8* |
|  | type | 1 |  | **Projected avg. row size** | | | | 10 |
|  | *Total* | *17* |  | **Avg. num rows** |  | | | 400 |
| **Projected avg. row size** | | 17 |  | **Size of table (data)** | |  | | 4000 |
| **Avg. num rows** |  | 8 |  | **Fields overhead** | |  | | 64 |
| **Size of table (data)** |  | 136 |  | **Index overhead** | |  | | 1600 |
| **Fields overhead** |  | 128 |  | **Projected table size** | |  | | **7363** |
| **Index overhead** |  | 36 |  |  | |  | |  |
| **Projected table size** |  | **360** |  |  | |  | |  |
|  |  |  |  |  | |  | |  |
| ScreeningItem |  |  |  | ScreeningProgramSamplers | | | | |
| **Fields** |  | Avg. field size (bytes) |  | **Fields** | | |  | Avg. field size (bytes) |
|  | program\_id | 4 |  |  | | | id | 4 |
|  | tag | 6 |  |  | | | tag | 6 |
|  | parameter\_id | 4 |  |  | | | status | 1 |
|  | *Total* | *14* |  |  | | | *Total* | *11* |
| **Projected avg. row size** | | 17.5 |  | **Projected avg. row size** | | | | 13.75 |
| **Avg. num rows** |  | 600 |  | **Avg. num rows** | | |  | 150 |
| **Size of table (data)** |  | 10500 |  | **Size of table (data)** | | |  | 2062.5 |
| **Fields overhead** |  | 96 |  | **Fields overhead** | | |  | 96 |
| **Index overhead** |  | 4200 |  | **Index overhead** | | |  | 750 |
| **Projected table size** |  | **19235** |  | **Projected table size** | | |  | **3781** |
|  |  |  |  |  | | |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ParameterThresholds | |  |  | SamplerMedia |  |  |
| **Fields** |  | Avg. field size (bytes) |  | **Fields** |  | Avg. field size (bytes) |
|  | name | 8 |  |  | name | 4 |
|  | parameter\_id | 4 |  |  | content | 3670016 |
|  | min | 8 |  |  | tag | 6 |
|  | max | 8 |  |  | *Total* | *3670026* |
|  | *Total* | *28* |  | **Projected avg. row size** | | 4587532.5 |
| **Projected avg. row size** | | 35 |  | **Avg. num rows** |  | 800 |
| **Avg. num rows** |  | 320 |  | **Size of table (data)** |  | 3670026000 |
| **Size of table (data)** |  | 11200 |  | **Fields overhead** |  | 96 |
| **Fields overhead** |  | 128 |  | **Index overhead** |  | 1600 |
| **Index overhead** |  | 1920 |  | **Projected table size** |  | **4771036005** |
| **Projected table size** |  | **17222** |  |  |  |  |
|  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ScreeningFrequency | | | |  |  | Waterbody | |  |  |
| **Fields** |  | | | Avg. field size (bytes) |  | **Fields** | |  | Avg. field size (bytes) |
|  | id | | | 4 |  |  | | name | 11 |
|  | description | | | 16 |  |  | | type | 1 |
|  | frequency | | | 7 |  |  | | total | 12 |
|  | tag | | | 6 |  | **Projected avg. row size** | | | 15 |
|  | *Total* | | | *33* |  | **Avg. num rows** | |  | *40* |
| **Projected avg. row size** | | | | 41.25 |  | **Size of table (data)** | |  | 600 |
| **Avg. num rows** | |  | | 100 |  | **Fields overhead** | |  | 64 |
| **Size of table (data)** | | |  | 4125 |  | **Index overhead** | |  | 220 |
| **Fields overhead** | | |  | 128 |  | **Projected table size** | |  | **1149** |
| **Index overhead** | | |  | 200 |  |  | |  |  |
| **Projected table size** | | |  | **5789** |  |  | |  |  |
|  | | |  |  |  |  | |  |  |
|  | | |  |  |  |  | |  |  |
|  | | |  |  |  |  | |  |  |
|  | | |  |  |  |  | |  |  |
|  | | |  |  |  |  | |  |  |
|  | | | |  |  |  |  | |  |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cooperate |  | | | |  | |  |  |  |  |
| **Fields** |  | | | | Avg. field size (bytes) | |  |  |  |  |
|  | id | | | | 11 | |  |  |  |  |
|  | contractor\_name | | | | 9 | |  |  |  |  |
|  | laboratory\_name | | | | 9 | |  |  |  |  |
|  | *Total* | | | | *12* | |  |  |  |  |
| **Projected avg. row size** | | | | | 12 | |  |  |  |  |
| **Avg. num rows** | | |  | | 35 | |  |  |  |  |
| **Size of table (data)** | | |  | | 420 | |  |  |  |  |
| **Fields overhead** | | |  | | 96 | |  |  |  |  |
| **Index overhead** | | |  | | 16 | |  |  |  |  |
| **Projected table size** | | |  | | **638.4** | |  |  |  |  |
|  | | |  | |  | |  |  |  |  |
|  | | |  | |  | |  |  |  |  |
| Total database size | |  | | 4985282412 | |
| Extra allocation | |  | | 997056240.6 | |
| **Projected database size (MB)** | | | | **5705.202** | |

## Miscellaneous Notes

From the calculations, it is evident that the majority of space usage is consumed by sampler media, which may potentially contain a number of large images and videos. This is also despite the large number of samples assumed.

Also, as this estimation assumes that no data will be archived and removed from the database, it is likely to be an over-estimate on the likely actual database size.

However, as storage hard disks are incredibly cheap today, even the cheapest of new hard disks should provide far more storage than required by the system. As this will be installed on its own dedicated server, space usage shouldn’t be a particular concern for the TiGERS system. Also note that large amounts of data won’t have any serious effects on performance of the system as most of this data is from the sampler media and indexes will ensure the samples table performs efficiently.

Note that this estimate does not include all indexes which may be added to the system, which could cause notable increases in size. This is difficult to predict at this stage, because optimisation is performed late in the SDLC.