Proposal for uploading timestamped observations

In this document we propose some extensions to the clinical data input into TranSMART. We like to add:

- The possibility to specify unit information for an observation.
- Specify a date-time point for an specific observation.
- Group observations. For example PSA values at different points in times.

To realize this we have to:

1. Extend the format of the input-files
2. Store this information into the database of TranSMART

There is of course also the point what to do with this "new" data in TranSMART, but that's another issue outside the scope of this document.

Format of the input data-files

How the input files for TranSMART should look like is explained in "Dataset Explorer ETL Guide" from Recombinant written on januari 31, 2012.
In short: you at least should have 2 files

- Data-file:
  A tab-seperated file in which each row contains information about 1 specific patient and each column expresses the value for a specific observation. One column must unique specify the subject.
- Column-map file:
  Give some additional information about the columns in the "data-file".

The interesting part here is the column-mapping file. This file gives some extra information of the columns in the "data-file". It is the definition of this file were we would like to propose some extensions (marked red).
The bases of the column-mapping file looks like:

Filename	Category Number	Column Number	Data Label	Data Label Source	Control Vocab Cd

- Filename:
  The name of the data file containing the observations
- Category Code:
  Also called the "Concept Code", this defines the path in the "Data Explorer" tree of TranSMART where the observation will be shown.
- Column Number:
  Gives the column number in the "data-file" where the values for this observation are given.
- Data Label:
  The label you want to use for this observation (the leaf in the Dataset Explorer). The following words are reserved and have special meaning:
  - OMIT : Skip this column (equal to not mentioning the column)
  - SUBJ_ID : This column contains the subject id's (mandetory)
  - SITE_ID : This column contains the id of the site the observations are coming from.
  - VISIT_NAME : This column specifies the name of the visit the observation were done.
  - DATA_LABEL : Treat the data in this column as a data label for another column.
  - \ : The column number for the specific observation can be found in column "Data Label Source'
  - TIMESTAMP : The timestamp belonging to a specific observation
  - MODIFIER : Group observations. Can be used to upload time-series or can be used as annotation to an observation (i2b2)
  - UNITS : The unit associated with a specific observation.
- Data Label Source:
  Depending on column "Data Label" it means:
  - \ : The column number where to find the "Data Label" for the specific observation.
  - TIMESTAMP : The column number in the data file this column (timestamp) belongs to.
  - MODIFIER : The column number, containing the "default" observation.
  - UNITS : The column number in the data file this column (units) belongs to.
- Control Vocab CD
  If "Data Label" is "MODIFIER" then the contents of this is used as the "modifier_cd" for the observations

With "MODIFIER" you can group simular observations together. The observation with the "real" Data Label (not MODIFIER) is the value taken when a single value from the group is requested (default value). Al other observations in the group (with MODIFIER as the DATA Label) point to this default value.

Let's use an example to make things more clear.....

Data file (test-data.tsv)

Subject	Age	Unit	Weight	Date	Unit	Weight	Date	Weight	Date
1	40	year	60	1/1/2010	kg	65	1/1/2005	70	1/1/2000
2	50	year	70	1/2/2010	kg	68	1/2/2005	72	1/2/2000
...									

A column map file could be

Filename	Category Number	Column Number	Data Label	Data Label Source	Control Vocab Cd
test-data.tsv		1	SUBJ_ID		
test-data.tsv	Subjects+Demographics	2	Age		
test-data.tsv	Subjects+Demographics	3	UNITS	2	
test-data.tsv	Subjects+Demographics	4	Weight		
test-data.tsv	Subjects+Demographics	5	TIMESTAMP	4	
test-data.tsv	Subjects+Demographics	6	UNITS	4	
test-data.tsv	Subjects+Demographics	7	MODIFIER	4	
test-data.tsv	Subjects+Demographics	8	TIMESTAMP	7	
test-data.tsv	Subjects+Demographics	6	UNITS	7	
test-data.tsv	Subjects+Demographics	9	MODIFIER	4	
test-data.tsv	Subjects+Demographics	10	TIMESTAMP	9	
test-data.tsv	Subjects+Demographics	6	UNITS	9	

Note: from the grouped observations for "Weight" (4 + 7 + 9) the values from column 4 will be taken as the default value.

Where to put it in the database.

The data from the input data files is a little transformed and then put into the landings zone in table "tm_lz.lt_src_clinical_data". We like to extend this table a little

```
CREATE TABLE tm_lz.lt_src_clinical_data (

    study_id     character varying(25),
    site_id      character varying(50,
    subject_id    character varying(20),
    visit_name   character varying(100),
    data_lable    character varying(500),
    data_value   character varying(500),
    units_cd       character varying(50),                - - new
    timestamp    timestamp without time zone,    - - new
    modifier_cd  character varying(100),              - - new
    category_cd character varying(250),
    ctrl_vocab_  character varying(200)
)
```

If we take the input files from above the table would be populated like:

study_id	site_id	subject_id	visit_name	data_label	data_value	units_cd	timestamp	modifier_cd	category_cd	ctrl_vocab
CSTEST		1		Age	40	year			Subject+Demographics	
CSTEST		2		Age	50	year			Subject+Demographics	
CSTEST		1		Weight	60	kg	1/1/2010	@	Subject+Demographics	
CSTEST		2		Weight	70	kg	1/2/2010	@	Subject+Demographics	
CSTEST		1		Weight	65	kg	1/1/2005	SERIES:1	Subject+Demographics	
CSTEST		2		Weight	68	kg	1/2/2005	SERIES:1	Subject+Demographics	
CSTEST		1		Weight	70	kg	1/1/2000	SERIES:2	Subject+Demographics	
CSTEST		2		Weight	72	kg	1/2/2000	SERIES:2	Subject+Demographics	

Note1: The default item in a group, get's the special "modifier_cd" value "@". You can think of it as "SERIES:0". This is the value to be used if a single value is expected. The numbering is following the order they have in the column mapping file.
Note2: In case the the "Control Vocab Cd" column is filled in the column mapping file, this value will be used as the "modifier_cd". Make sure the modifier_cd (Control Vocab Cd) is unique within a grouped grouped observations.

Data from the table "tm_lz.lt_src_clinical_data" will be put in the "i2b2" tables by the stored procedure "i2b2_load_clinical_data". The new columns "units_cd", "timestamp" and "modifier_cd" will go into already existing items "units_cd", "start_date" and "modifier_cd" of the table "i2b2demodata.observation_fact" respectively.

We now have a way to fill some extra database fields, which were already availabele in the used "i2b2" tables. Now we only have to do something useful with this extra information 🙂.