LINKÖPINGS UNIVERSITET

GMS Data Analysis

João Pedro
Johannes Rajala
2007

Table of contents:

1 Introduction

The project described here was developed on the course of Information Visualization (TNM048) [4] as its Final Project. The developed application provides the user to visually explore and analyze musical preferences and their relations to social, economical and geographical data within Europe. The application was designed taking into account human perception and interaction issues to provide easy access to users with different backgrounds.

The document is presented using the Sense-Making loop (Figure 1), which describes the steps of an analytical reasoning process. /2/

The application was developed using Geo-Analytics Visualization Framework, shortly GAV, a framework developed at VITA Group/LiU. GAV provides developers with basic visualization tools for tailor-made and task-oriented applications development. This project extends GAV with components for multi-window Treemap, an embedded TableLens and descriptive glyphs. /4/

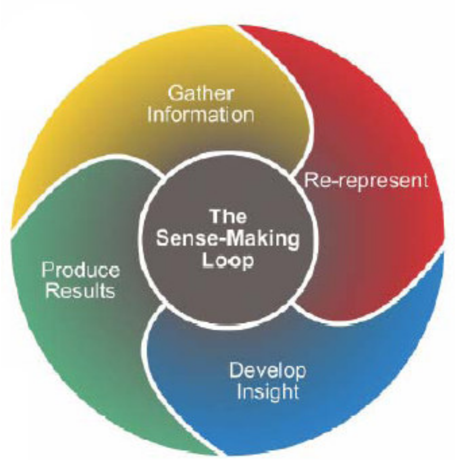


Figure 1 The Sense-Making loop of analytical reasoning process

2 Reasoning Process

The reasoning process was followed quite strictly. The followed approach was directed to an exploratory analysis of the data, given the fact that no assumptions existed about it; though, some confirmatory analysis has also been done. After the information has been collected, the visual representation of the data has been thoroughly analysed with usability studies and personal decisions. The choices produced a good insight of the data, which lead to also good results.

2.1 Gathering Information

The gathered information was collected from three different sources and was crossed using custom-built Data Miner.

2.1.1 Data Sources

The three sources of data had some shared attributes, making them, therefore, good for Data Mining. It should be noticed, however, that the data contained on the musical databases is not on a standardized format. This issue is discussed deeply on the Data Mining sub-section.

2.1.1.1 FreeDB.org

This database contains nearly 22 Million albums from the whole world. The extracted features from this database where:

- Album Name
- Artist Name
- Music Style Name

The database was downloaded completely – around 600 MB. The information was stored in ASCII format, in which each album was represented by a single file and stored in a hierarchical way – 11 folder representing global styles of music, e.g.: rock, pop, etc.

2.1.1.2 MusicBrainz.org

The MusicBrainz database is much smaller than the FreeDB database, but it contains important features to be crossed with latter. This database is also downloadable, but it requires the installation of a SQL server, and all the scripts to create and set it up are done in Linux. Therefore, the decision of using a C# Client was made. The client was embedded in out Data Mining application, and information was fetched using Internet.

The data attributes fetched from this database where:

- Albums releases years
- Release Countries
- Artists

2.1.1.3 CIA World Factbook

The CIA World Factbook is a huge collection of data available online in HTML format – usually in form of tables. It is completely up to date and it contains lots of relevant information per Country.

It has been decided that just a small group of attributes would be collected to join the other collected data, since the Data Mining must be done manually – copy, pasting and formatting in Excel – and is just a part of the process of creating the whole application.

The chosen attributes where:

- List of Countries
- Per Country
    - Government Type
    - Median Age
    - GDP per Capita
    - Unemployment Rate

2.1.2 Data Mining

2.1.2.1 *Manual* Data Parsing

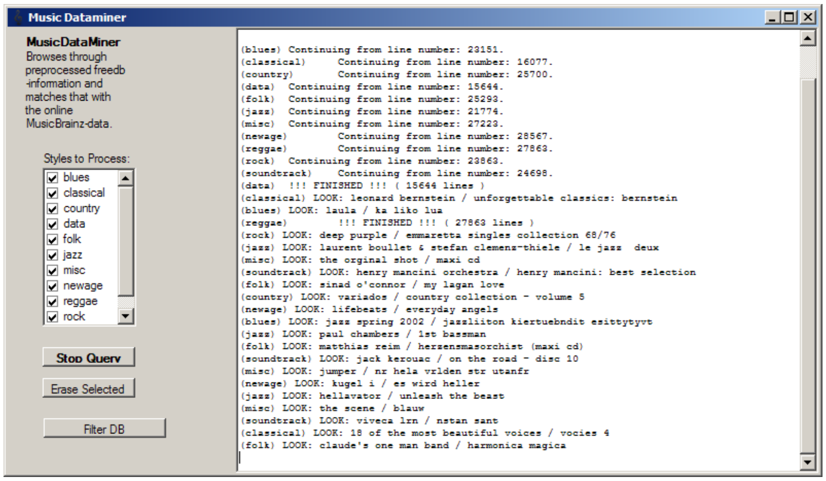
Some data, as referred before, was mined manually due to limitations on format and other issues. The data collected from the CIA's World Factbook database was copied from HTML to Excel format. Once in Excel, the data was added as column attributes per country.

The other database that required manual parsing was FreeDB. Some files contained albums written in Unicode format, but the MusicBrainz's database did not support that format. For each of the 11 music styles, an output file was generated with the result of parsing all the input files on that folder (mentioned in the previous section) and skipping the ones containing only Unicode characters.

2.1.2.2 Musical DataMiner Application

The Musical DataMiner application (Figure 2) was created to fetch the data from the MusicBrainz online server, cross that data with FreeDB and World Factbook databases and produce a final, serializable result in a format convenient (common data structures like Hashtables and Lists) for further usage on the final application.

The application was developed for Multithreaded usage and to be able to stop and restart on the previously stop point. It also contains a feature to filter data for European countries only – the reasons for that are explained further on the document.



MusicDataMiner
Browses through
preprocessed freedb
-information and
matches that with
the online
MusicBrainz-data.

Styles to Process:
☑ blues
☑ classical
☑ country
☑ data
☑ folk
☑ jazz
☑ misc
☑ newage
☑ reggae
☑ rock

```
(blues) Continuing from line number: 23151.
(classical)     Continuing from line number: 16077.
(country)       Continuing from line number: 25700.
(data)  Continuing from line number: 15644.
(folk)  Continuing from line number: 25293.
(jazz)  Continuing from line number: 21774.
(misc)  Continuing from line number: 27223.
(newage)        Continuing from line number: 28567.
(reggae)        Continuing from line number: 27863.
(rock)  Continuing from line number: 23863.
(soundtrack)    Continuing from line number: 24698.
(data)  !!! FINISHED !!! ( 15644 lines )
(classical) LOOK: leonard bernstein / unforgettable classics: bernstein
(blues) LOOK: laula / ka liko lua
(reggae)        !!! FINISHED !!! ( 27863 lines )
(rock) LOOK: deep purple / emmaretta singles collection 68/76
(jazz) LOOK: laurent boullet & stefan clemens-thiele / le jazz  deux
(misc) LOOK: the orginal shot / maxi cd
(soundtrack) LOOK: henry mancini orchestra / henry mancini: best selection
(folk) LOOK: sinad o'connor / my lagan love
(country) LOOK: variados / country collection - volume 5
(newage) LOOK: lifebeats / everyday angels
(blues) LOOK: jazz spring 2002 / jazzliiton kiertuebndit esittytyvt
(jazz) LOOK: paul chambers / 1st bassman
(folk) LOOK: matthias reim / herzensmasorchist (maxi cd)
(soundtrack) LOOK: jack kerouac / on the road - disc 10
(misc) LOOK: jumper / nr hela vrlden str utanfr
(newage) LOOK: kugel i / es wird heller
(jazz) LOOK: hellavator / unleash the beast
(misc) LOOK: the scene / blauw
(soundtrack) LOOK: viveca lrn / nstan sant
(classical) LOOK: 18 of the most beautiful voices / vocies 4
(folk) LOOK: claude's one man band / harmonica magica
```

Figure 2 - Music DataMiner Application

2.2 Re-Representation

2.2.1 Usability Study

With the gathered data, it is possible to draw some conclusions, but in order to have an insight to the data proper visualizations have to be chosen. Already in an early phase, simple qualitative user studies were performed in order to obtain information in the selection of:

- Colours
- Interfaces and Visual Representation
- Correct and Easy interaction

These studies were helpful and provided new views and ideas to visualize even more items that are interesting.

One of the problems in visualization was the large amounts of nominal data. Such nominal attributes as:

- Country names
- Albums and Artists
- Government Types
- Music styles

Many of the attributes can be considered as ordinal data and while it was possible to show these on parallel coordinates, the treemap was considered more informative when comparing this kind of ordinal data.

Usability study was done by eight people and consisted of quantitative and qualitative part. (See appendix 1 and 2). All views were considered useful and tasks were performed in a reasonable amount of time (mostly in less than minute). The piechart below shows how the Treemap is almost as fast to use as the Parallel coordinates.

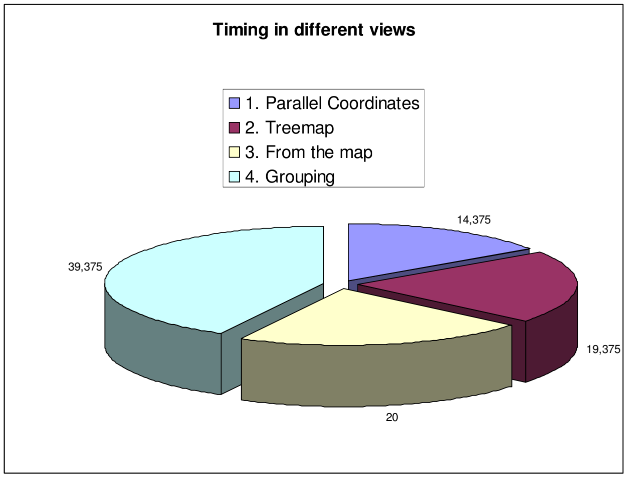


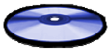
Figure 3 - Quantitative measurements of usability study

2.2.2 Presentation

For the visualizations the three connected views, Geographical Map, Treemap, and Parallel Coordinates with a simple table lens to increase perception, was observed to be optimal.

Geographical Map
As the gathered information was strongly connected to the countries of the world, using a geographical map provided very informative and quick to adapt to interface even for the non-experienced users.

Glyphs
Here instead of using generic types of glyphs, such as pie charts or Chernoff faces, users seemed to prefer the customized descriptive glyphs instead. The glyphs were stacked on top of each other, and the highest pile represents highest values in that attribute.

			
Albums	GDP	Employment	Median Age

Colours
The colour is linked with the other views, and can be selected to show either different attributes or groups. Selection is shown by showing thicker border around a country and increasing the intensity of the selected region.

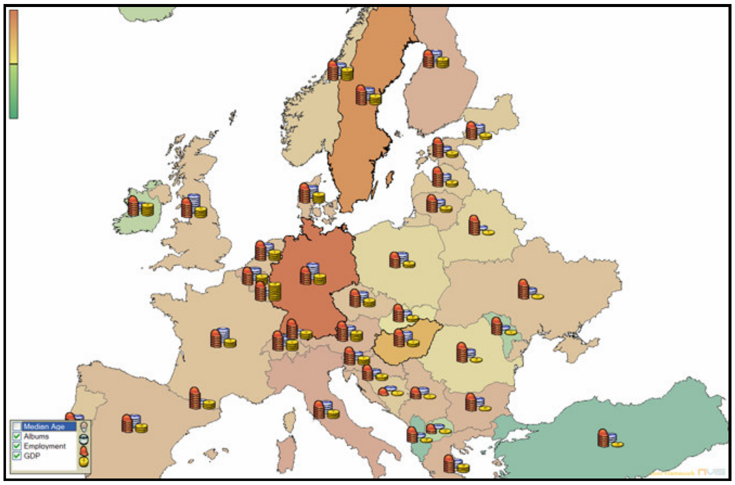


Figure 4 - Geographical Map

Treemap

The amount of nominal data was apparent when the data gathering was on going, but a lot of it could also be considered ordinal data. This allowed many different kinds of grouping. For representing ordinal groupings, the Treemap provided very good visualization approach. From different implementations, the Squarified Treepmaps seemed the most appropriate and visually pleasing approach. /1/

The Treemap is constructed from three levels:

- Root level
- Ordinal level for grouping
- Quantitative level for calculating suitable areas for all the levels

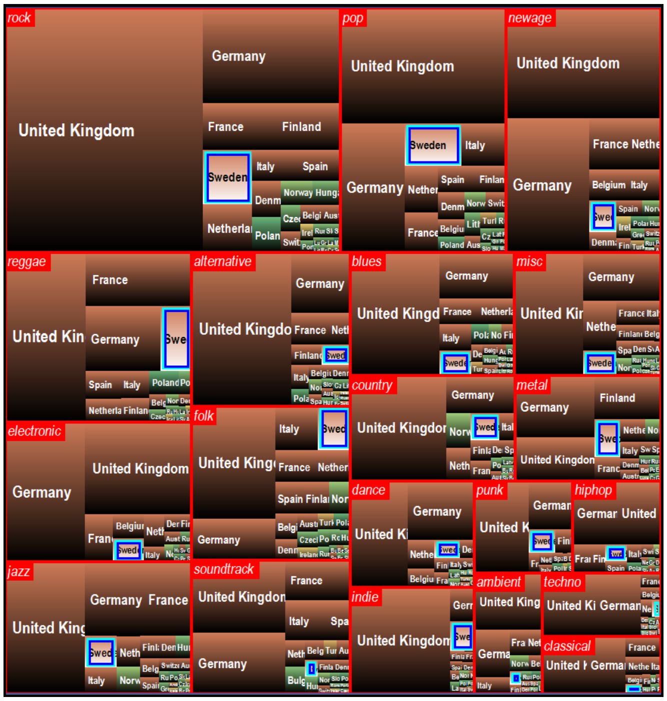
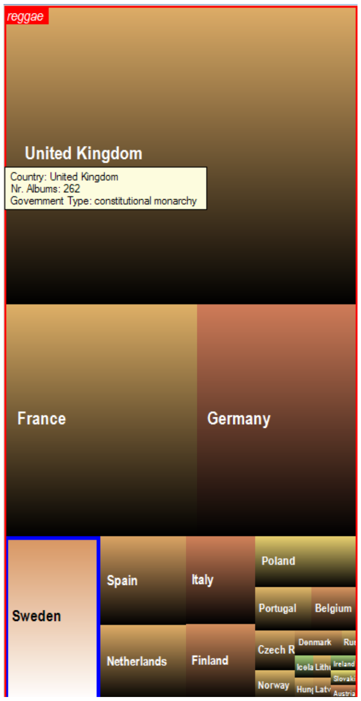


Figure 5 - Treemap with Styles and Countries as ordinal and amount of Albums as Quantitative data



Country: United Kingdom
Nr. Albums: 262
Government Type: constitutional monarchy

Figure 6 - Treemap Zoomed in (Reggae Style), with Tooltip being shown

Various kinds of groupings can be done using Treemap, and the application supports the following:

- Styles per country
- Countries per Style
- Countries per Government

Colouring

The red colour proved to be easiest to distinguish groups. The basic colours are the same than in other views but are drawn with a gradient to black for distinguishing the Treemap cells, while optimizing the screen space. For selection the blue and cyan borders help to locate items in the Treemap.

Parallel Coordinates

Finding trends, locating outliers, and to be able to freely compare attributes, especially quantitative, the Parallel Coordinates was an ideal solution. As the data was having not only many attributes but also lot of rows, for which it was necessary to display nominal information, it was needed to have a way to show data even in a restricted screen space. For this, a simple table lens proved to be a suitable solution. The Parallel Coordinates in its basic form worked nicely otherwise.

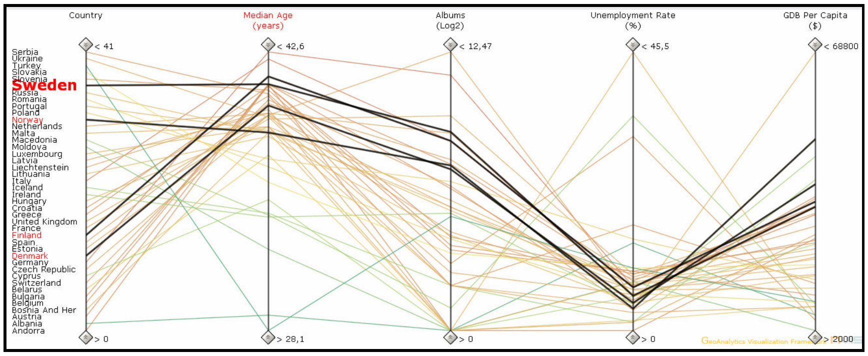


Figure 7 - Parallel Coordinates and a simple Table Lens

Colouring
Colours of the lines can be set according to the attributes in the parallel coordinates. Visual queries can be made with the axis filter, and picking can be used to locate countries interesting attributes. The selection works and shows also on the Table lens.

2.3 Insight

The layout of the application was chosen to allow the user to gain a deeper insight of the data through easy manipulation of its graphical interface. The usability studies were very useful to manage issues like colour, mouse interaction, size, location and many other important features.

2.3.1 Application Layout

The application layout is divided into three main views (with a supplementary one):

- Treemap View
- Parallel Coordinates View
- Map Plot View

As said before, most of our data is Nominal, which constraints the list of visual representations to just a few, justifying the above choices.
All three components are synchronized by a common ColorMap, which makes the comprehension of data much easier. That ColorMap is created based on any of the Parallel Coordinates attributes, and can be controlled by a colour legend locate on the top left corner of the Map Plot. Those colours where chosen in order to be pre-attentively distinguished and without tiring the eyes. Therefore, the choice has fallen on a pastel palette.

2.3.1.1 Main Treemap (Countries per Music Style)

The main Treemap displays countries grouped by music style. On this view, it is easy to see, for a given style, the countries that have the biggest amount of album releases in it. The zoom function allows the user to dig up into the style and eventually find countries with less releases. The colour map helps on linking a country on the Treemap with its representation in other view.

2.3.1.2 Map Plot

The Map Plot is probably the most intuitive component. Any person used to geography, can quickly locate a country, select it, and see it as well selected on other views. It is good to distinguish features on regions (See *Results*).

The glyphs were chosen to be as more intuitive as they can be:

- Median Age: represented by an old man face. More faces depict higher Median Age on that country.
- GDP per Capita: represented by a dollar gold coin (since the GDP units are in dollars). The gold colour, the coin shape and dollar sign, all represent money;
- Unemployment Rate: represented by a worker's orange helmet. More helmets depict less employment, thus more unemployment; the orange colour was used to distinguish from the blue of the CD's and coin's golden yellow.
- Albums: represented by a CD. More CD's mean more released albums on that country; blue was used instead of silver, since it might be confused (according to usability studies) with coins silver colour.

The user also can also choose the number of used glyphs, but the usability studies have given two (2) as being the best number without visual clutter and for more clearance. Therefore, the default is set to two.

2.3.1.3 Parallel Coordinates

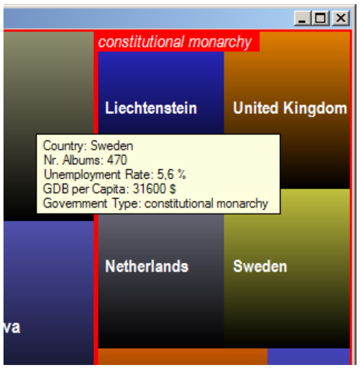
The Parallel Coordinates Plot is the complementary view of the main three. It provides the textual basis to locate the countries, as mentioned previously. The additional feature added is the ability to group or cluster the data, using the K-Means algorithm. This, as it will be mentioned further, was useful for the quick detection of trends and outliers.

2.3.1.4 Additional Treemaps

Together with the main Treemap, three more Treemaps are available as mentioned before (one is the same as the main one – described ahead). Those representations (see example in Figure 10) are separated in another window because their colouring is not related to the colouring of the data represented on the main view. There is, though, one exception. The main Treemap (countries per style) is replicated here so that the user can be able to compare two different styles at the same time. Regarding the other Treemaps, the colouring has been chosen to be quite different so the user cannot, unconsciously, relate any of those with the main view.

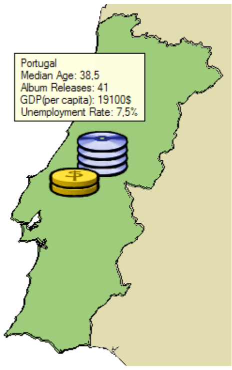
2.3.1.5 Brushing

The information displayed on all the previously mentioned views is not sufficient, though, to give the user full information about the data. *Brushing*, is a very useful and intuitive way of providing the user with easy, although not pre-attentive, information. All the views, except the Parallel Coordinates, have Tooltip information. See Figure 8 and 9 for reference.



Country: Sweden
Nr. Albums: 470
Unemployment Rate: 5,6 %
GDB per Capita: 31600 $
Government Type: constitutional monarchy

Figure 8 - Brushing in Government Types Treemap



Portugal
Median Age: 38,5
Album Releases: 41
GDP(per capita): 19100$
Unemployment Rate: 7,5%

Figure 9 - Brushing in MapPlot

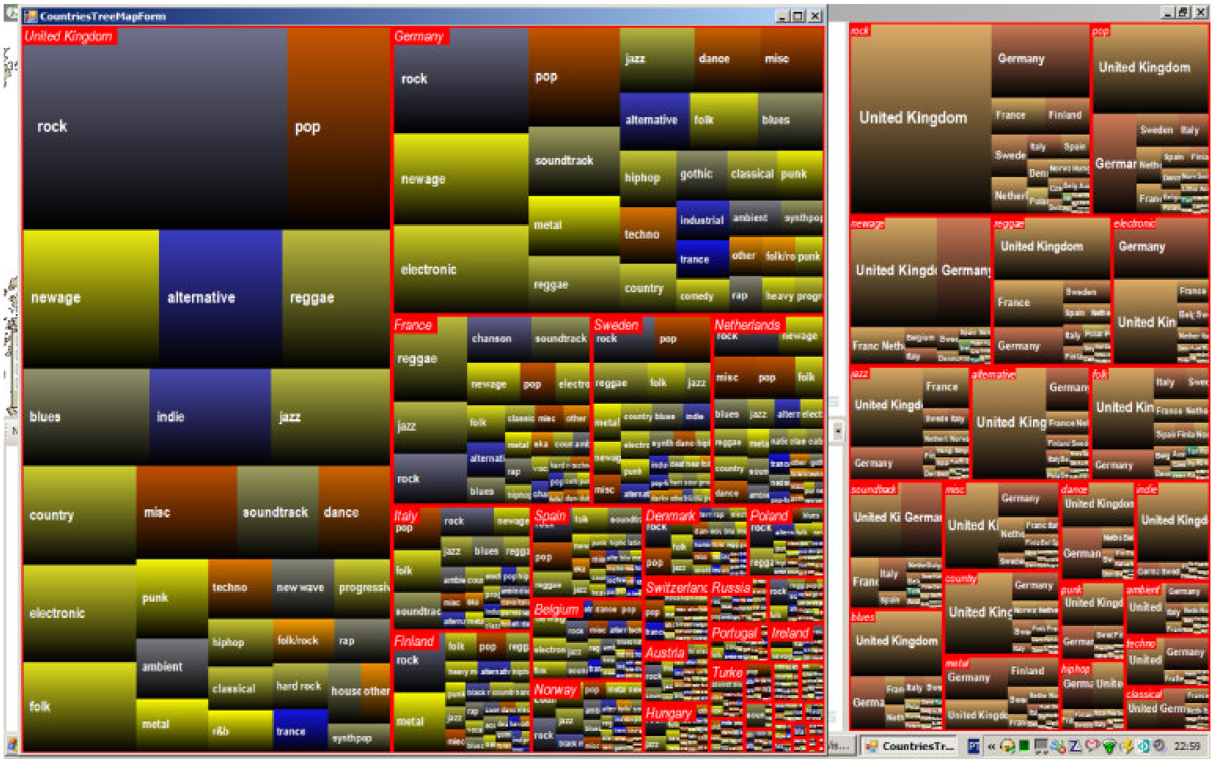


Figure 10 - Comparing different data using Main and Additional Treemaps

3 Results

With the methods previously described it was possible to get very good insight into the musical trends inside the Europe. The purpose was to compare data from countries all over the world, but the used framework had a limitation of not supporting the map format of the world. As geographical data was important part of using users' cognitive understanding, it was decided to filter our data to only show and compare the attributes inside Europe.

In the project, the data mining had a very important role as non-standardized data was being used, and also because three databases were crossed. In some cases this may have biased the dataset, as some items could not be synchronized between the databases. This was because of non-uniform specifications and interfaces between the data sources.

The application was tested mostly in exploratory sense, but also as confirmatory, as some assumptions were made already in the data mining phase and even before that.

Here are some of the gained insights:

- Scandinavian Countries have very similar habits in social data and in the music data.
- Eastern Europe is strongly connected to a single group in the data we gathered, but this could be because of their music releases are written in un recognized character set, because people have less access on the Internet.
- A Geographical trend of three eastern European countries was located by using the grouping (Seemingly the unemployment rate is very different from the neighbouring countries):
    - Bosnia and Herzegovina
    - Serbia
    - Macedonia
- Luxemburg seems to be a very strong social outlier, with a very high GDP and very low Unemployment

This application was interesting to use for many people in user test as they were interested seeing how their home country was performing against other countries, but also to have an insight what music is released in their country.

4 Conclusions

The project provided a good understanding of musical influences in Europe, but also how to provide usable visualizations of otherwise complicated and interconnected data.

During the project, it became also evident that datamining is an important part of the project and has to be well thought out.

Usability study was important already in the early design of the interface to find appropriate colouring scheme and easy-to-use user interface.

5 References

1. Bruls M., Huizing K, Van Wijk J.J., Squarified Treemaps, 1999, Eindhoven, Netherlands.
2. Jern, Mikael, Information Visualization Notes, 2007, Norrköping, Sweden.
3. Panopticon Website, Treemaps, http://www.panopticon.com, 2007.
4. VITA/LiU, GAV Framework. http://vita.itn.liu.se, 2007-03-30.

6 Appendices

1. Usability Study Questions
2. Usability Study Results

1 Usability Study Questions

GMS Visualizer – Usability Study

With GM application you can compare geographical, musical, and sociological data. The data has been gathered from three different sources and linked with appropriate attributes.

freedb.org	- Album, artist, Style
musicbrainz.org	- Album release country, year
CIA world factbook	- Government type, Median age, GDP per capita, unemployment

QUESTIONS:

1. **From the parallel coordinates:**
   Locate A country with highest ***Median Age*** by selecting it

   ____

2. **From the Treemap**
   Given the country you selected, find a style where it has released more albums than other countries.

   ____

3. **From the Map**
   Choose ***Median Age*** glyph and find the country with lowest median age.

   ____

4. **Group countries into two groups**
   Which of the countries in the ***less releasing group*** has most releases in "**HipHop" Style.**

   ____

5. **How Did you experience the Navigation**

   ____

6. **Did the Colors help you?**

   ____

   ____

7. **How many glyphs are usable at the same time in your opinion?**

2 Appendix: Usability Study Results

Finding information	**User1**	**User2**	**User3**	**User4**	**User5**	**User6**	**User7**	**User8**	**Avg**
0. Experience level	high	high	high	low	high	low	low	low	
1. Parallel Coordinates	5	10	10	15	10	30	5	30	14,4
2. Treemap	15	20	20	15	15	20	20	30	19,4
3. From the map	20	15	25	15	15	30	20	20	20
4. Grouping	20	45	60	40	35	45	20	50	39,4
5. Feelings about the navigation	positive	neutral	positive	positive	neutral	neutral	positive	negative	++
6. Color Usage	positive	negative	positive	neutral	neutral	positive	positive	neutral	++++
7. Max amount of glyphs for comparisons	2	4	4	2	2	2	2	4	2,75