

ECP 2005 CULT 038097/Bernstein

BERNSTEIN

Deliverable no. 15, ref. D4.4 Statistical functionalities

Deliverable number	<i>D4.4</i>
Dissemination level	<i>Public</i>
Delivery date	<i>30 April 2008</i>
Status	<i>Final</i>
Author(s)	<i>The Bernstein Consortium</i>



eContentplus

This project is funded under the *eContentplus* programme¹,
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

¹ OJ L 79, 24.3.2005, p. 1.

The Bernstein Project

Deliverable D 4.4

Statistical functionalities

Table of Contents

1	INTRODUCTION.....	3
2	REPLY FORMATS TO USERS' REQUESTS	3
3	NAVIGATION ON THE SEARCH PAGES	4
4	FUNCTIONALITIES ON THE STATISTICS PAGE	4
4.1	SUMMARY STATISTICS SECTION	4
4.2	CHOICE STATISTICS SECTION.....	7
4.3	DETAILED STATISTICS FOR SINGLE CRITERIA	8
4.4	DETAILED STATISTICS FOR PAIRED CRITERIA	14
5	HANDLING OF DATES.....	21

1 Introduction

With over 120.000 items in the combined databases of the project it becomes necessary to provide means to visualize the statistical properties of the conglomerate of paper watermarks. Indeed statistics are as important for historians and experts as information on individual items. At present, except a limited functionality in the WZMA database, none of the existing resources gives statistics of the holdings.

The statistical functionality is one of the two fundamental ways in which users can apprehend the data on paper history made available through Bernstein: either at the object level of individual watermark descriptions, or at group level of selected watermarks according to the user's criteria. A quantitative description of the user's selection provides an insight into the structure of the data and allows its interpretation.

The task consists in implementing numerical and graphical statistics in the integrated workspace which will be responsible of gathering raw information on quantities of data from all the databases and process the data in such ways as to represent it in an unified form, given the differences in content and format. Statistical information will be provided in a numerical form and through powerful diagram visualizations.

Statistical capabilities are provided for “Simple Search” and “Advanced Search” results. Statistics are only accessible through the simple and advanced search pages - there will be no “statistics” page on the website menu.

In the following chapters the statistical functionalities of the integrated workspace are described.

2 Reply formats to users' requests

There are three formats for representing the reply of the Bernstein workspace to a user's request:

- (1) list of objects in the integrated databases matching the request's criteria
- (2) numerical and graphical statistics on the result set
- (3) cartographic representation of the reply

For each request made, users can select successively any number of these representations. Additionally users should have the option to download the statistical data for further processing and correlation with own data.

3 Navigation on the search pages

The details of the steps involved in the request/reply process are as follows:

- ❖ The user selects search criteria, either through the simple search input box or through the input boxes and options of the advanced search.
- ❖ If the total number of hits is below the limitation of hits, the user can select the mode for displaying the results (if the total number of hits exceeds the limitation of hits the user has to refine the query or increase the limitation of hits):
 - “List” for list of items
 - “Statistics” for numerical and graphical statistics
 - “Map” for cartographic representation
- ❖ If the chosen display mode is “list” or “statistics”, the results are displayed on the same page. For “map” new pages are generated.

4 Functionalities on the statistics page

The statistics page has three elements: a summary, a choice section and a detailed statistics section.

4.1 Summary statistics section

The summary gives basic statistics for each searchable and common criterion for the user's data subset.

<i>Criterion</i>	<i>Summary statistics</i>
Motif:	<ul style="list-style-type: none"> • quantity of types in selection (absolute values) • total number of types in the databases • percentage of types in selection out of the total number (relative values) • diagram (*)
Place of Use:	<ul style="list-style-type: none"> • quantity (of types in selection) • total number • percentage • diagram (*)
Depository:	<ul style="list-style-type: none"> • quantity (of types in selection) • total number • percentage • diagram (*)
Date:	<ul style="list-style-type: none"> • quantity (of types in selection) • first date • last date • arithmetic mean • standard deviation • diagram (**)
Height:	<ul style="list-style-type: none"> • quantity (of types in selection) • minimum • maximum • arithmetic mean • standard deviation • diagram (**)

Distance of Chainlines:	<ul style="list-style-type: none"> • quantity (of types in selection) • minimum • maximum • arithmetic mean • standard deviation • diagram (**)
-------------------------	---

(*) For this diagram a pie chart will be generated because this allows visualizing also very small percentage values (e.g. 0.10%) and therefore provides a better visual representation of the numbers than a horizontal box would offer.

(**) For this diagram a bar chart will be generated which shows the position of the (arithmetic) mean in regard to the minimum and maximum.

Note: If also the country is available for 'Place of Use' or 'Depository' in the databases then 'country' will be an additional criterion within the summary, choice and details section!

In the following example (“Example 1”) we assume the user searched for watermarks with date “1330” and selected as database only POL. The first reply informs the user that there are 44 hits in total and she decides to see the statistics. After selecting the radio button “Statistics” the user clicks to “Show Results” and gets the summary statistics (see Figure 1).

German [Français](#) [Contact](#) [Credits](#) Search Other Ressources

Watermark Catalog

BERNSTEIN  THE MEMORY OF PAPERS

Simple Search | Advanced Search | Component Search | Browse Motif

Databases

WZMA

NIKI

WILC

POL

TOTAL 44 100%

Hits

Motif

Place of Use

Depository

Date

Height

Distance of Chainline

Single Values

Place of Use

Depository

Date

Height

Distance of Chainlines

Combination-Matrix for statistical functions

<input type="checkbox"/>				
<input type="checkbox"/>				
<input type="checkbox"/>				
<input type="checkbox"/>				
<input type="checkbox"/>				
<input type="checkbox"/>				

continuous update the result

List Statistics Map

Date >> YYYY

[Add General Search](#) || [Add DB Specific Search](#) || [Help](#)

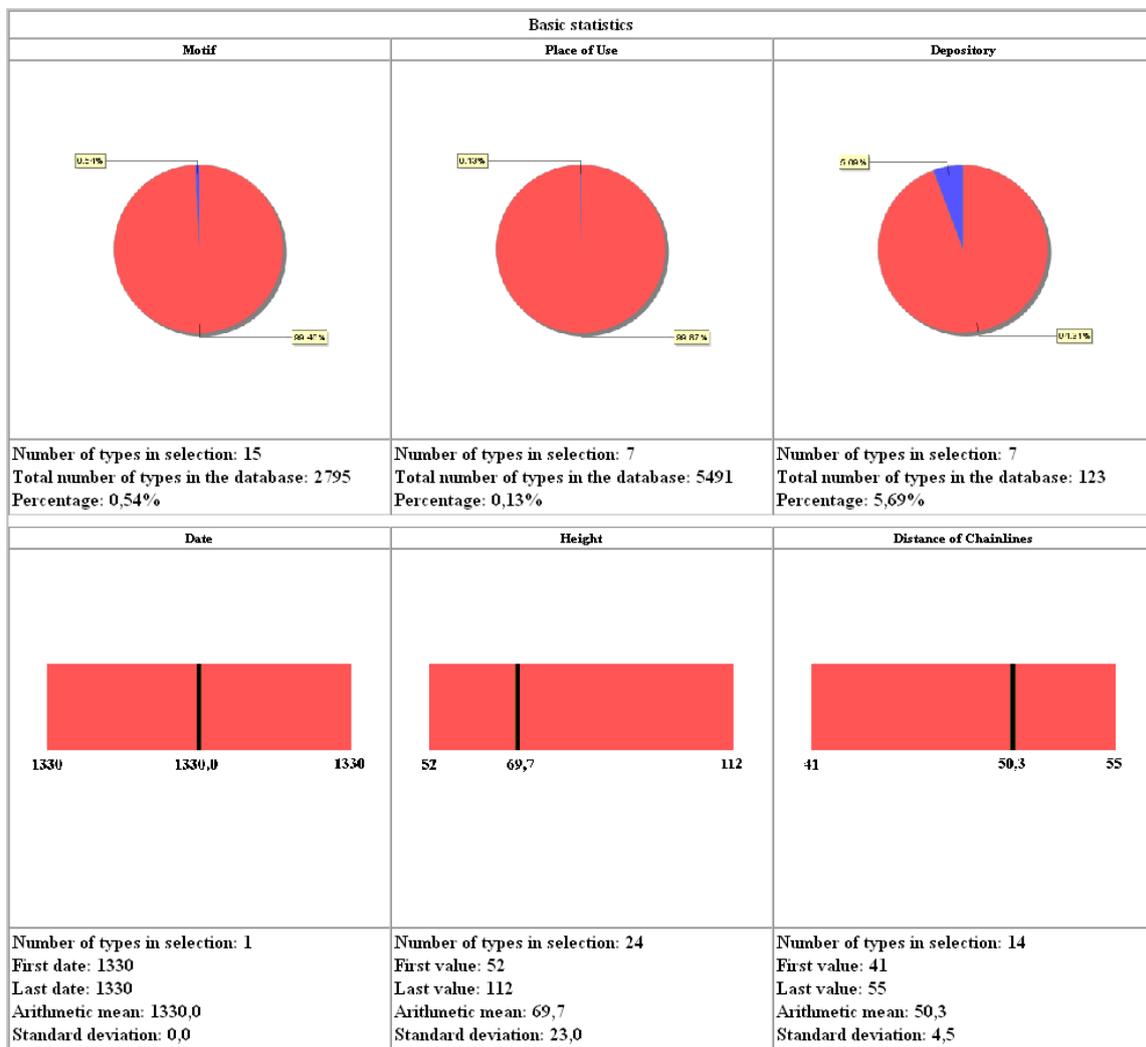


Figure 1: Summary statistics

4.2 Choice statistics section

The choice section allows the user to get more detailed statistics on results. It consists of radio buttons for selection of single and paired criteria.

		Single values	Paired values					
			Motif	Place	Depository	Date	Height	Distance
1.	Motif	X		X	X	X	X	X
2.	Place of Use	X			X	X	X	X
3.	Depository	X				X	-	-
4.	Date	X					X	X
5.	Height	X						-
6.	Distance of Chainlines	X						

Table 1: Choice statistics section

Detailed statistics for motif, place, depository, date, height or distance are generated only if there is more than one type in the selection.

For all detailed statistics there exists an upper limit for the number of types in the selection. If this limit is exceeded no single chart is generated but the detailed statistics is broken down by individual types and for each of them a separate chart is generated.

4.3 Detailed statistics for single criteria

If the user makes a choice in the detailed statistics section for single criteria, the results are displayed on the same page and any previously generated statistics will be overwritten.

For date, height and distance the detailed statistics will be extended with values of interquartile mean, range, standard deviation, skewness and kurtosis of the selected data.

An “Export” button allows the user to download the numerical data for further usages.

The following charts will be generated for the “Example 1” described above (search in POL for date “1330”).

1. Only Motif:

In the chart (see Figure 2) the number of occurrences of each motif is represented by the horizontal bar (e.g. the motif “Bull's head - With eyes, nose and further face features - Without additional motif” is the most frequently motif of the selection).

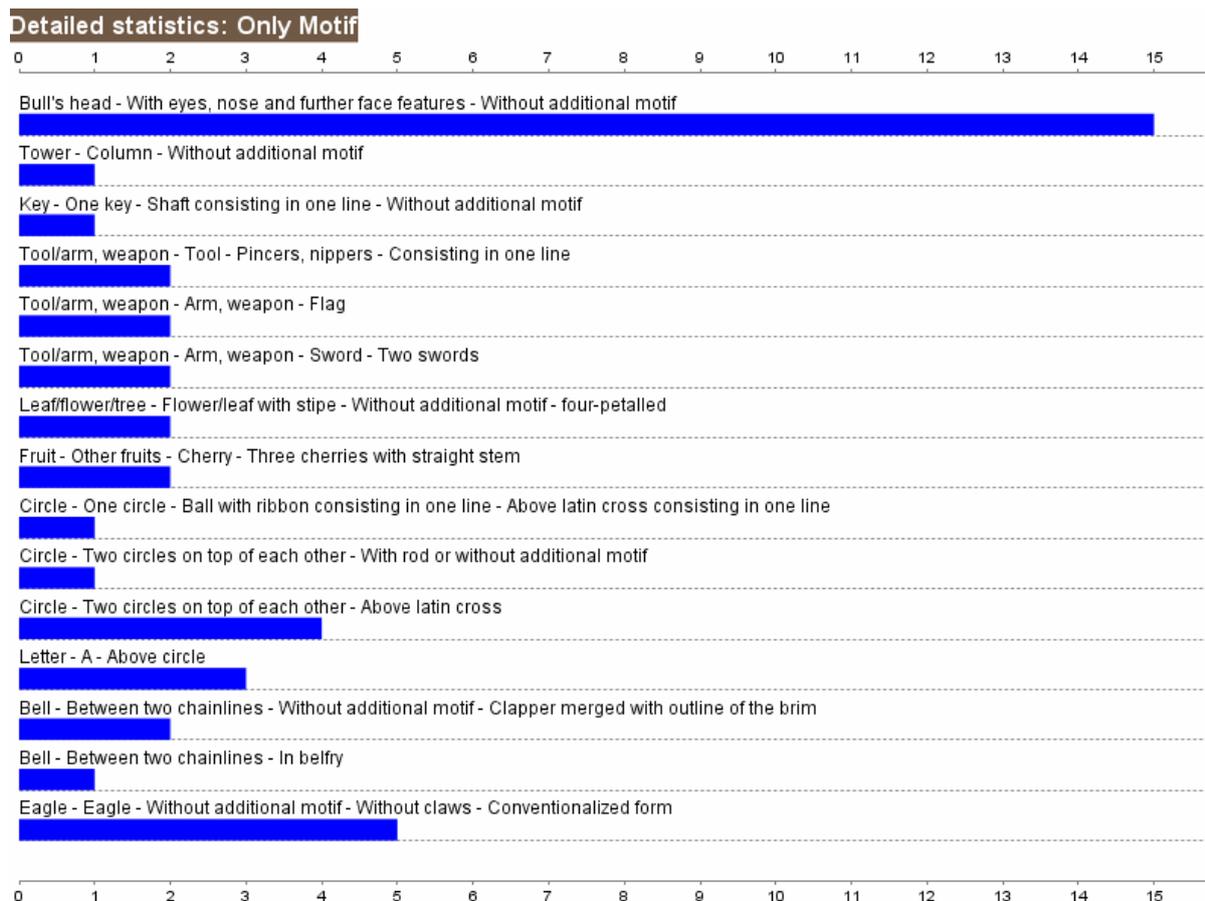


Figure 2: Detailed statistics: Only Motif

2. Only Place of Use:

In the chart (see Figure 3) the number of occurrences of each place is represented by the horizontal bar (e.g. the place “Lucca” can be found 8 times within the selection).

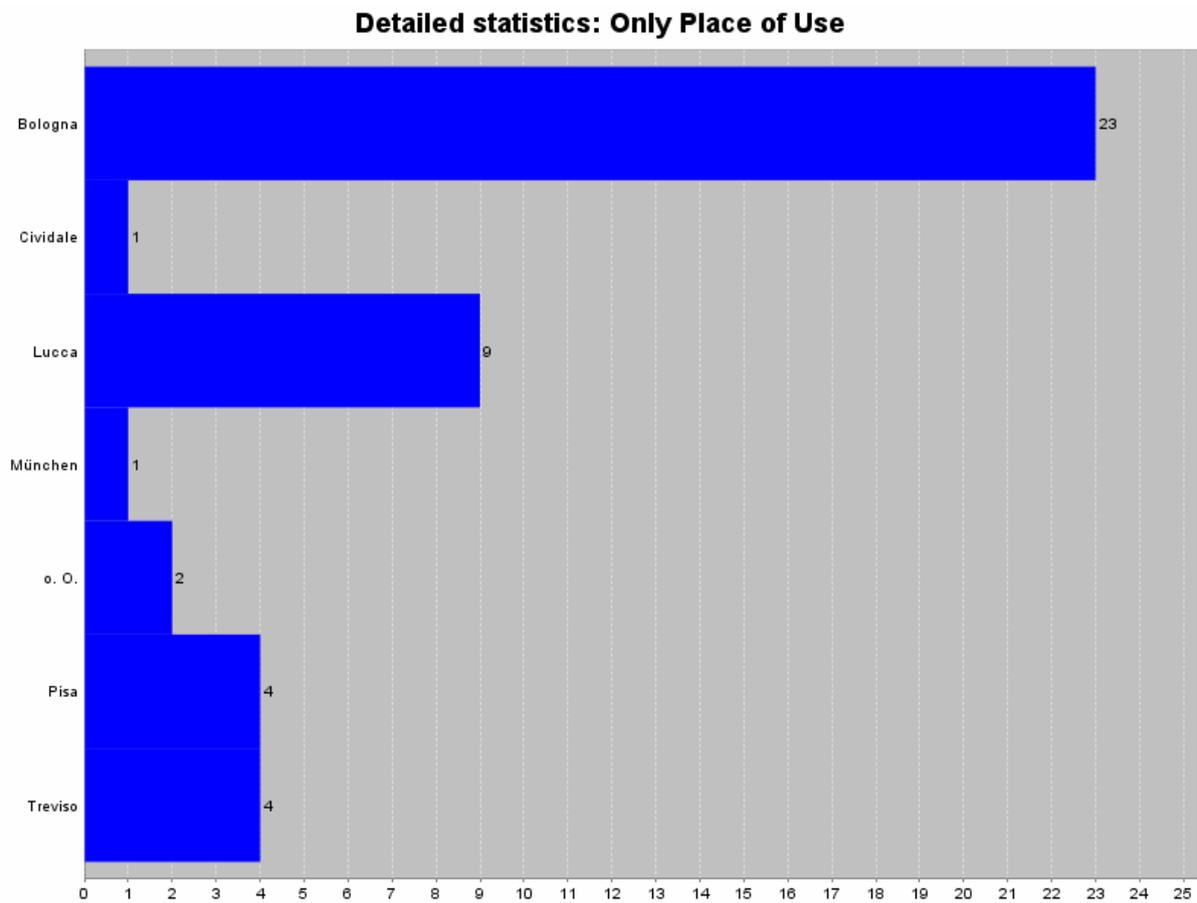


Figure 3: Detailed statistics: Only Place of Use

3. Only Depository:

Similar to the chart above the number of occurrences of each depository is represented by the horizontal bar (e.g. the depository “StA Treviso” can be found 4 times within the selection).

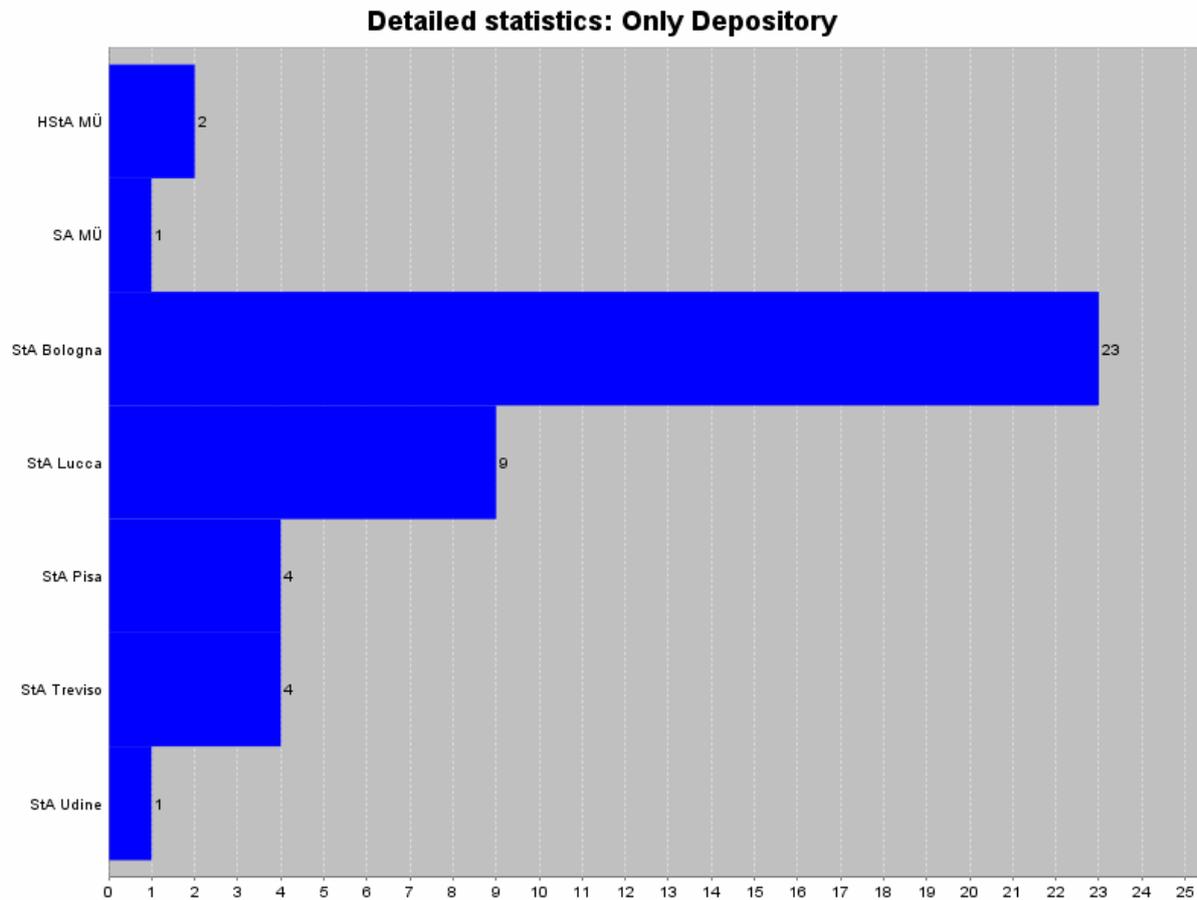


Figure 4: Detailed statistics: Only Depository

4. Only Date:

In the following example (“Example 2”) we assume the user searched for watermarks with motif “bird crown” and selected POL, WILC and WZMA as databases. The user gets 122 hits in total and decides to see the detailed statistics for date.

In the graphical histogram (see Figure 5) the weighted number of occurrences (see chapter 5) of each date is represented by the vertical bar (e.g. the year 1491 has a weight of 5.25).

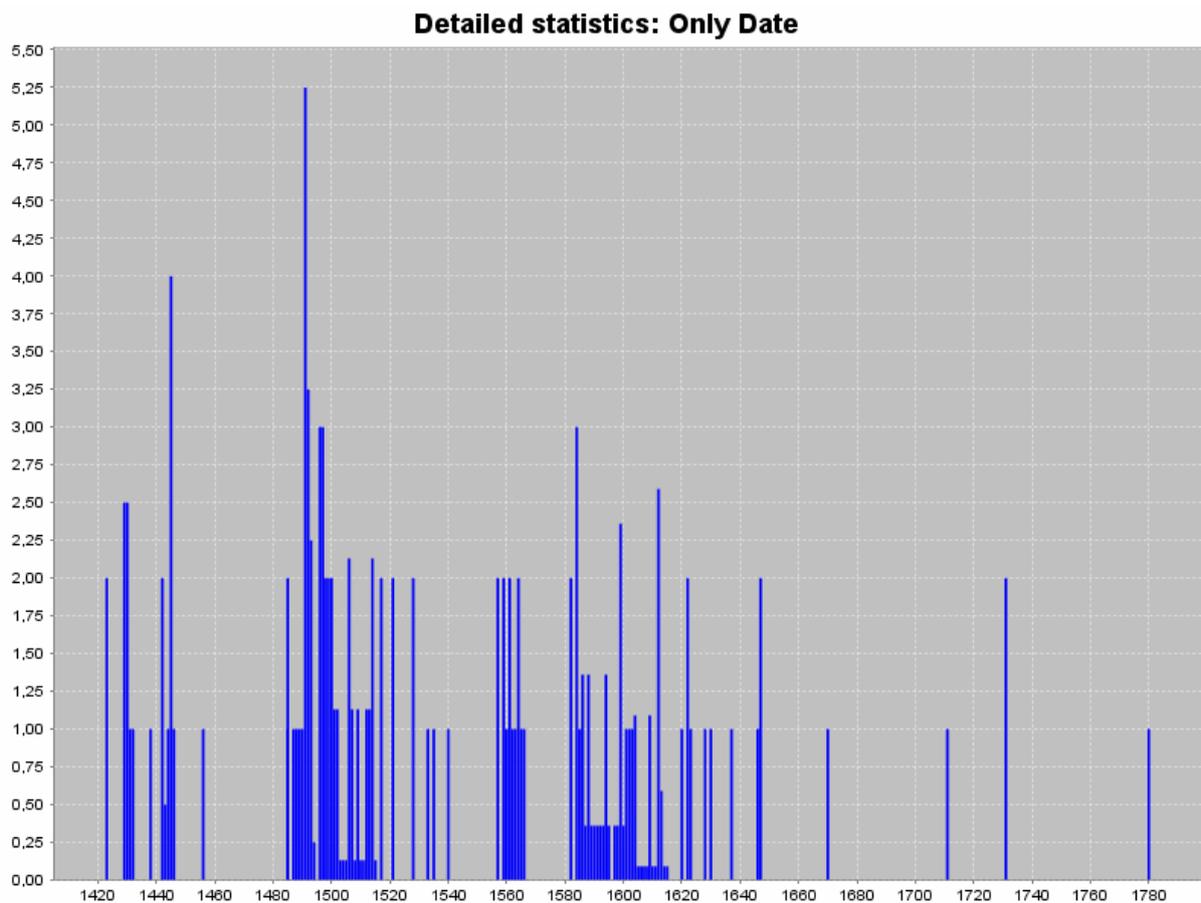


Figure 5: Detailed statistics: Only Date

The charts which follow are generated for “Example 1” (search in POL for date “1330”).

5. Only Height:

In the chart (see Figure 6) the number of occurrences of each height is represented by the vertical bar (e.g. the height “55” can be found 3 times within the selection and there is no height “85” within the result set).

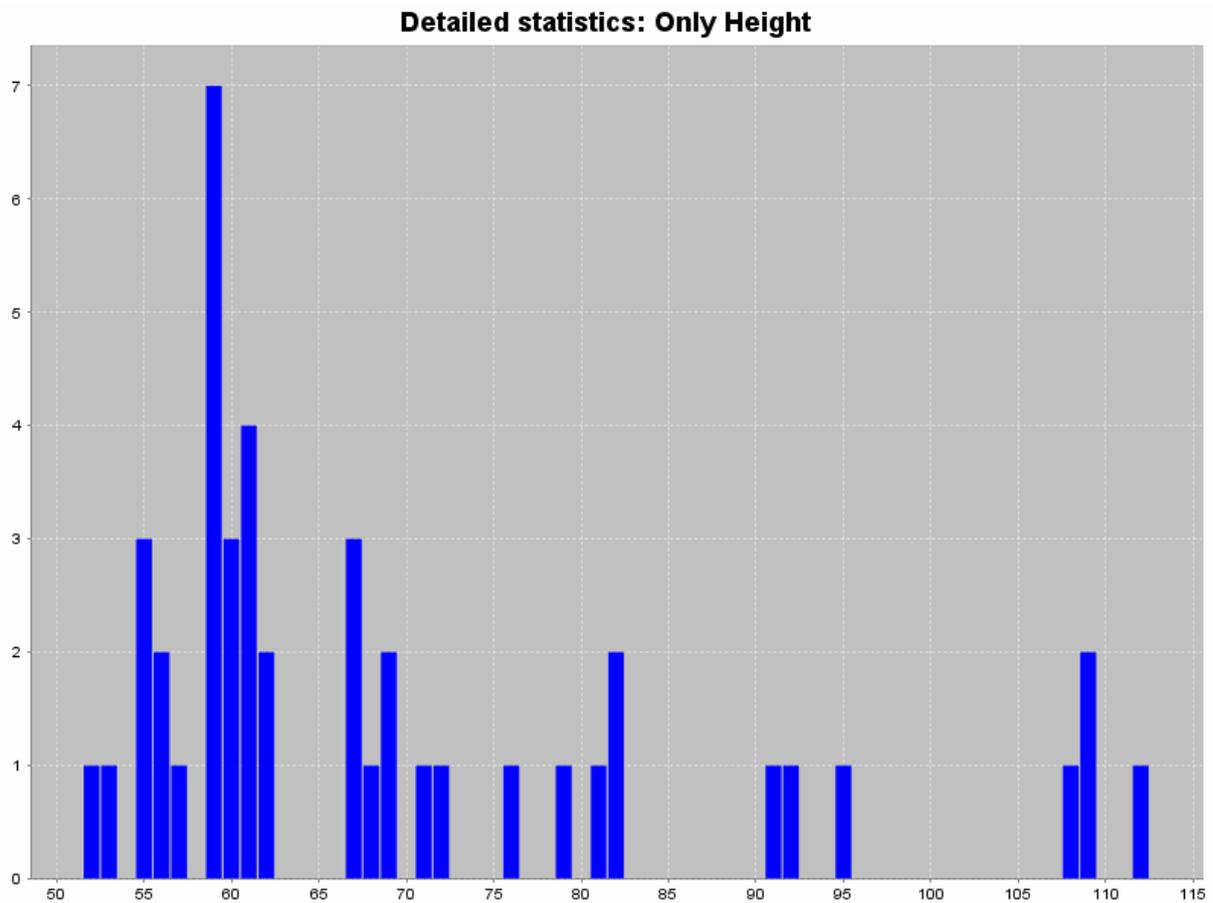


Figure 6: Detailed statistics: Only Height

6. Only Distance:

Similar to the chart above the number of occurrences of each distance of chainlines is represented by the vertical bar (e.g. the distance “49” can be found 4 times within the selection and there is no distance “46” within the result set).

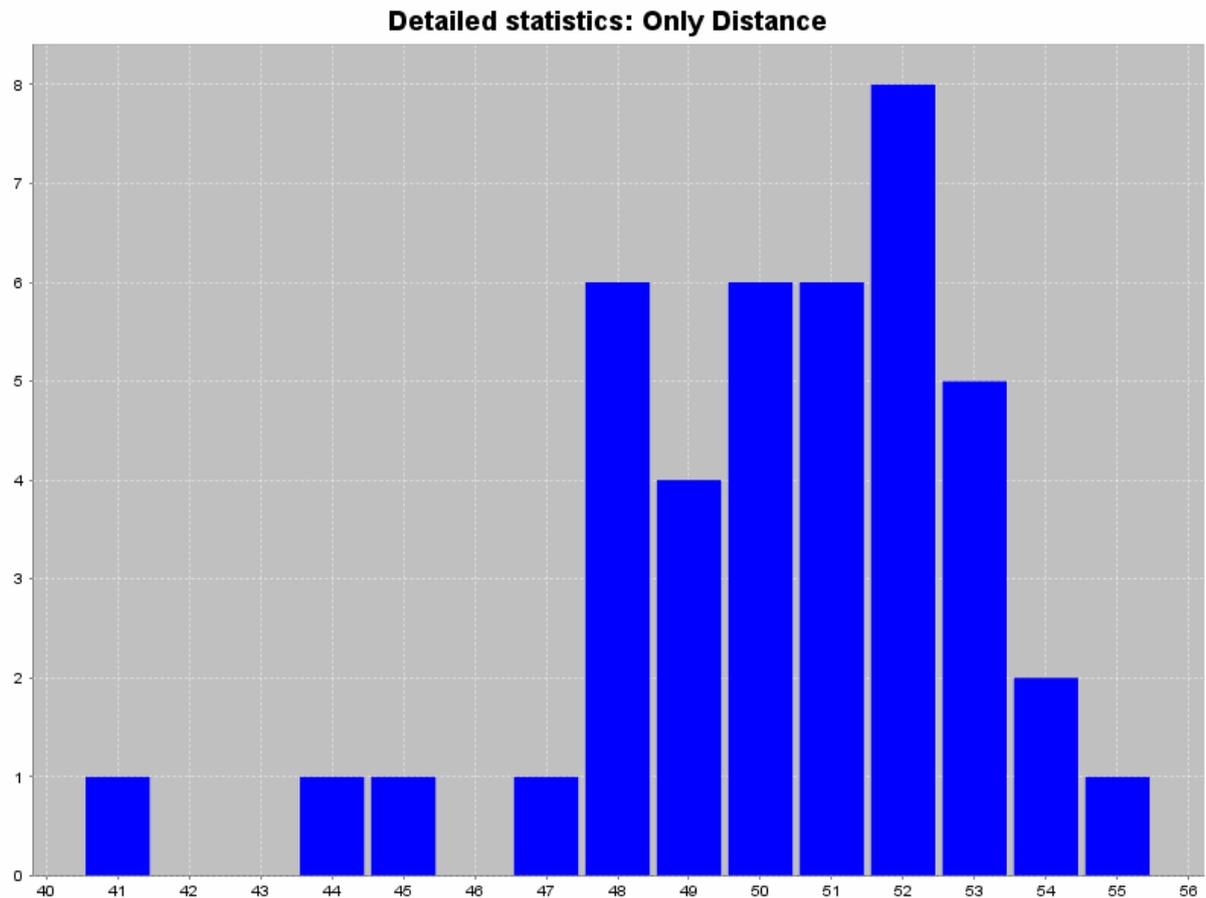


Figure 7: Detailed statistics: Only Distance

4.4 Detailed statistics for paired criteria

If the user makes a choice in the summary statistics section for paired criteria, the results are displayed on the same page.

If the user makes a paired selection between two statistical criteria, the results are displayed on the same page in a diagram.

1. Motif & Place of Use:

In the “Bubble” chart (see Figure 8) the 15 different motifs of the result set can be found on the x-axis and the places are represented by 7 different colours. (e.g. the motif “Bull's head - With eyes, nose and further face features - Without additional motif” [1] can be found most frequently in “Bologna” [dark blue] while in “Pisa” [magenta] you can find only the motifs “Circle - Two circles on top of each other - Above latin cross” [11] and “Bull's head - With eyes, nose and further face features - Without additional motif” [1]).

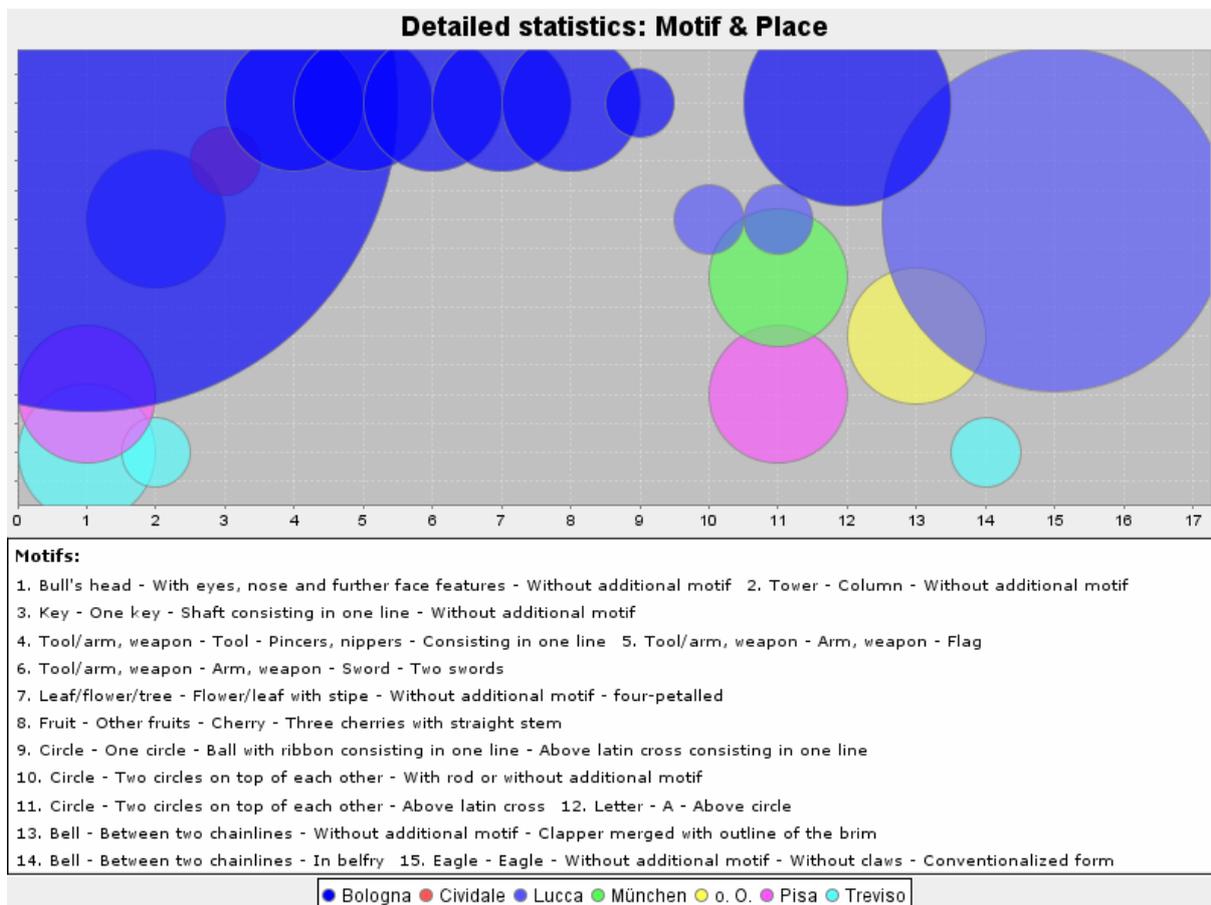


Figure 8: Detailed statistics: Motif & Place of Use

2. Motif & Depository:

The diagram for the “Motif & Depository” detailed statistics will be a “Bubble” chart similar to Figure 8.

3. Motif & Date

In the following example (“Example 2”) we assume the user searched for watermarks with motif “bird crown” and selected POL, WILC and WZMA as databases.

In the “MinMax“ chart (see Figure 9) the different motifs can be found on the x-axis and the first, interquartile mean and last date values for each motif are drawn on the y-axis (e.g. the dates for the motif “Bird - Eagle - Two heads - Above crown” are between “1512” and “1780”).

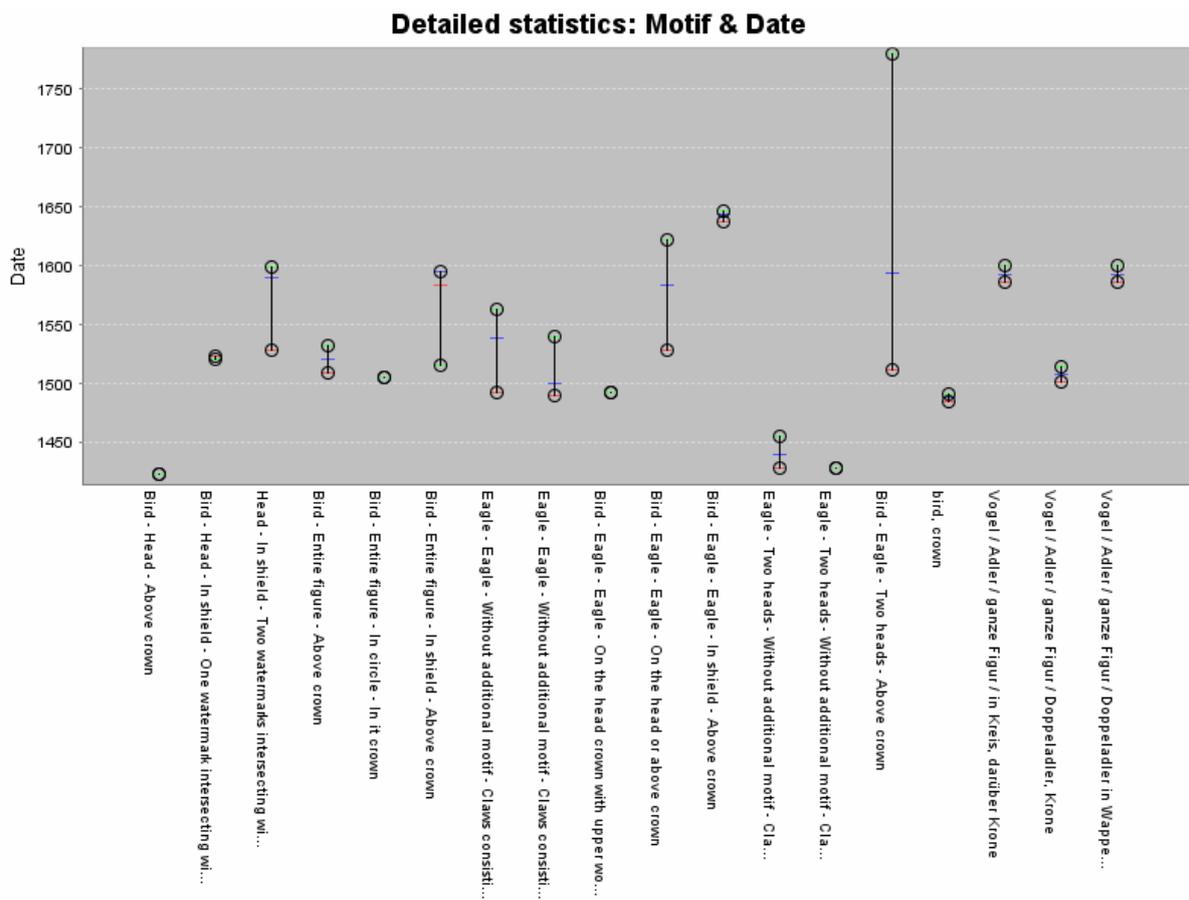


Figure 9: Detailed statistics: Motif & Date

The charts which follow are generated for “Example 1” (search in POL for date “1330”).

4. Motif & Height

In the “MinMax“ chart (see Figure 10) the 15 different motifs of the selection can be found on the x-axis and the minimum, interquartile mean and maximum height values for each motif are drawn on the y-axis (e.g. the height values for the motif “Circle - Two circles on top of each other - Above latin cross” are between “68” and “109” – the interquartile mean is about “84.2”).

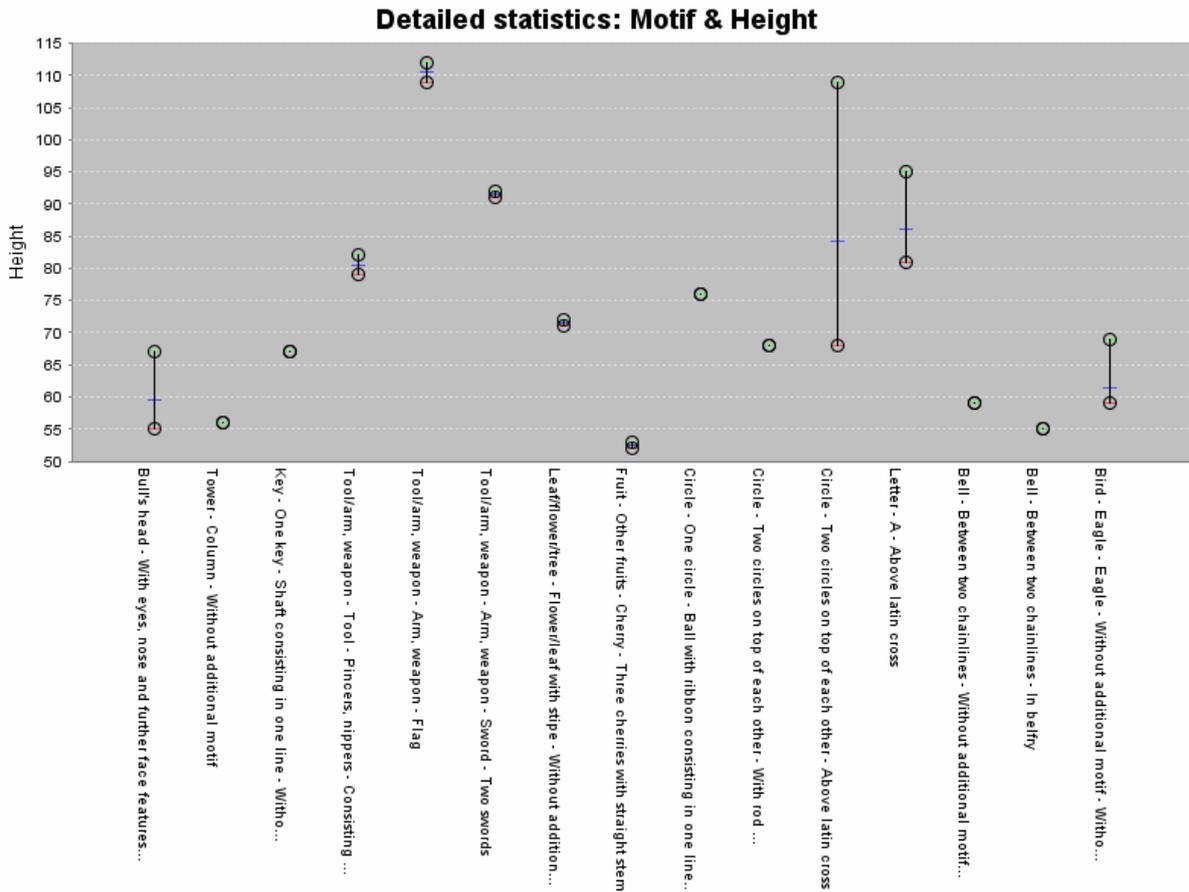


Figure 10: Detailed statistics: Motif & Height

5. Motif & Distance

In the “MinMax” chart (see Figure 11) the different motifs can be found on the x-axis and the minimum, interquartile mean and maximum distance values for each motif are drawn on the y-axis (e.g. the distance values for the motif “Circle – Two circles on top of each other – Above latin cross” are between “47” and “54”).

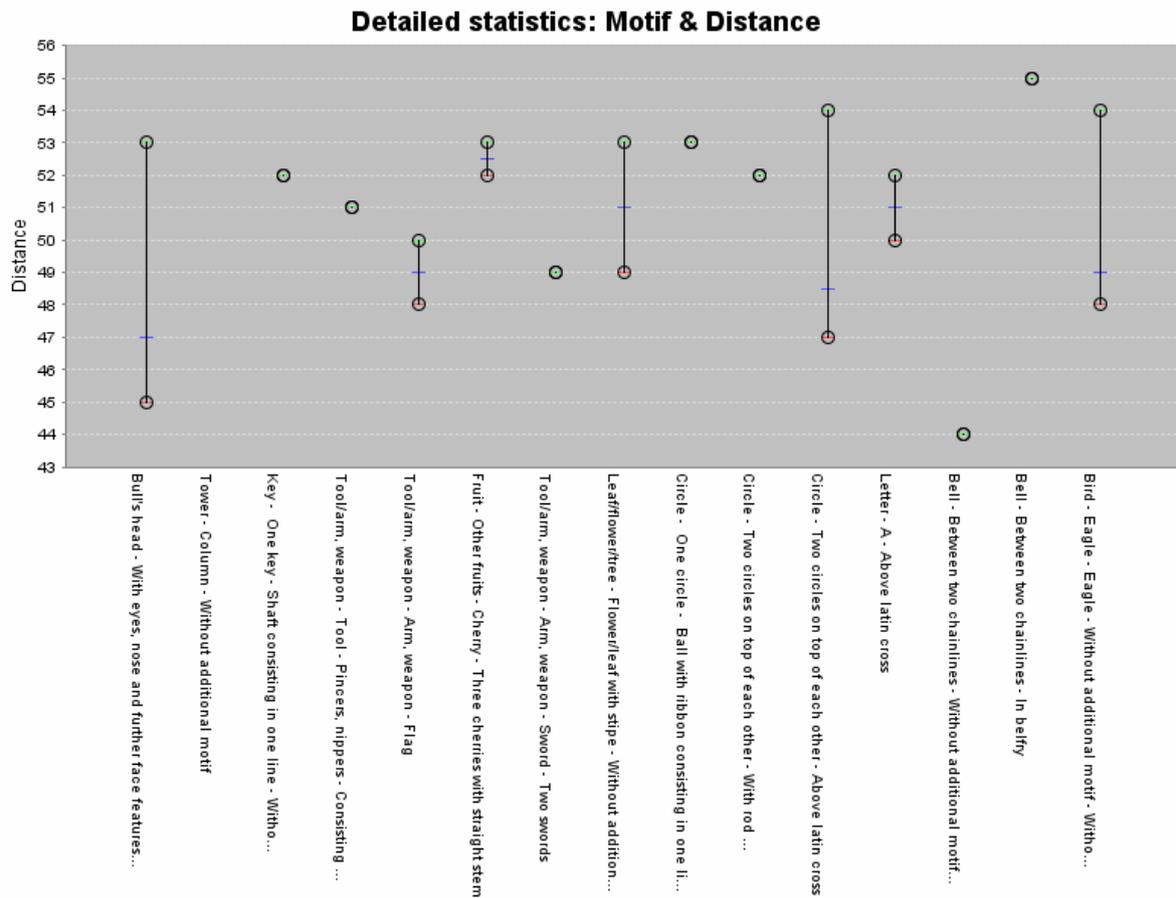


Figure 11: Detailed statistics: Motif & Distance

6. Place & Depository

The diagram for the “Place & Depository” detailed statistics will be a “Bubble” chart similar to Figure 8.

7. Place & Date

The diagram for these detailed statistics will be a “MinMax” chart analogous to Figure 11.

8. Place & Height

In the “MinMax“ chart (see Figure 12) the different places of the selection can be found on the x-axis and the minimum, interquartile mean and maximum height values for each of the places are drawn on the y-axis (e.g. the height values for the place “Bologna” are between “52” and “112” – the interquartile mean is about “70”).

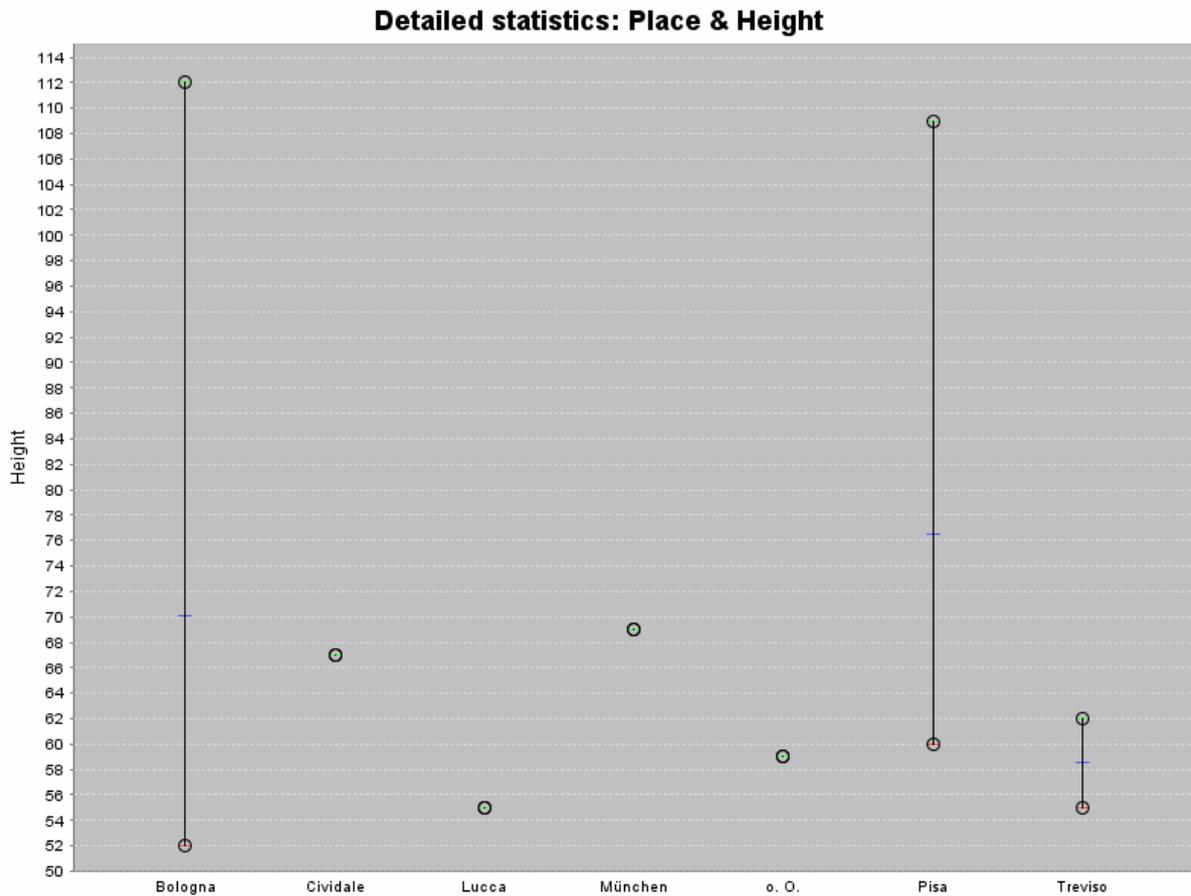


Figure 12: Detailed statistics: Place & Height

9. Place & Distance

In the “MinMax“ chart (see Figure 13) the 7 different places of the result set can be found on the x-axis and the minimum, interquartile mean and maximum distance values for each of the places are drawn on the y-axis (e.g. the distance values for the place “Pisa” are between “47” and “51” – the interquartile mean is about “48.8”).

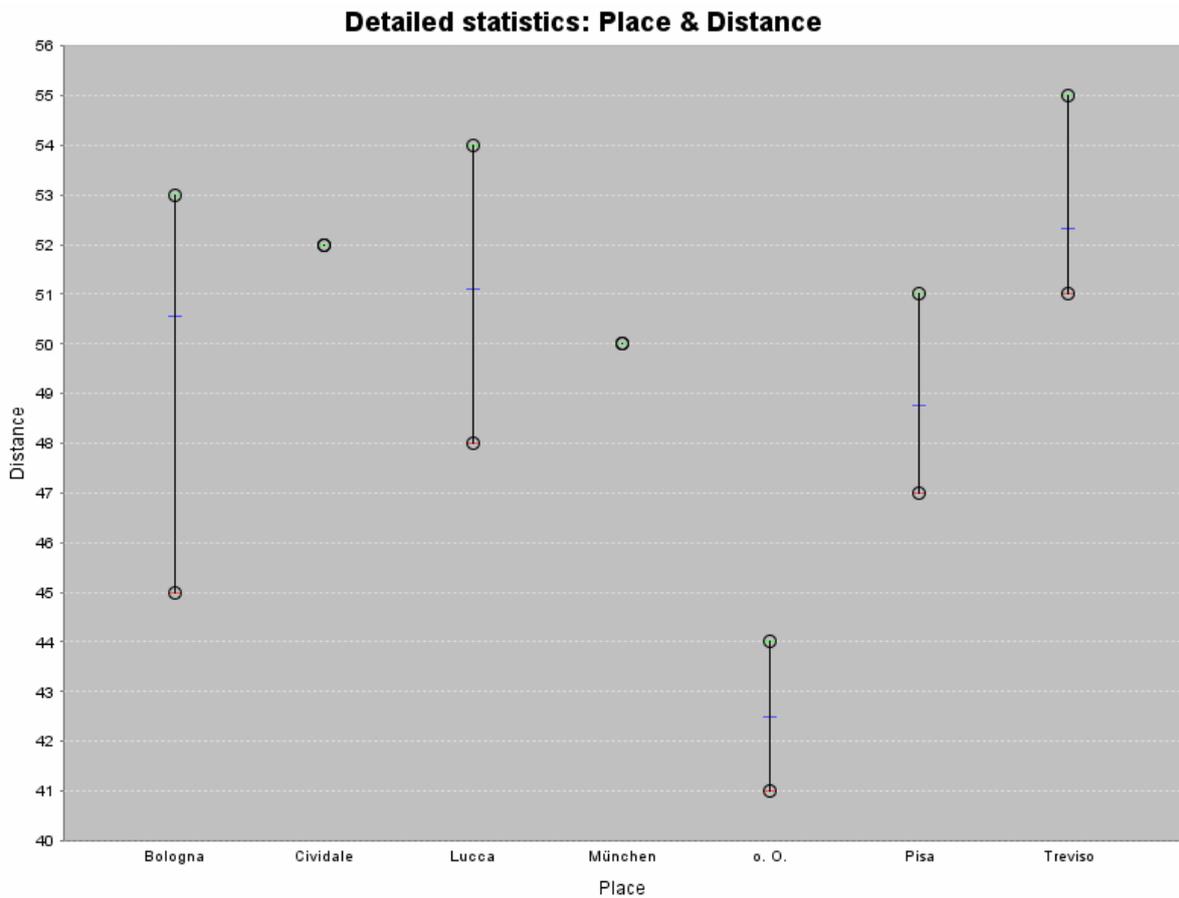


Figure 13: Detailed statistics: Place & Distance

10. Depository & Date

In the following example (“Example 2”) we assume the user searched for watermarks with motif “bird crown” and selected POL, WILC and WZMA as databases.

In the “MinMax” chart (see Figure 14) the 29 different depositories can be found on the x-axis and the first, interquartile mean and last date values for each depository are drawn on the y-axis (e.g. the dates for the depository “HStA” are between “1564” and “1609”).

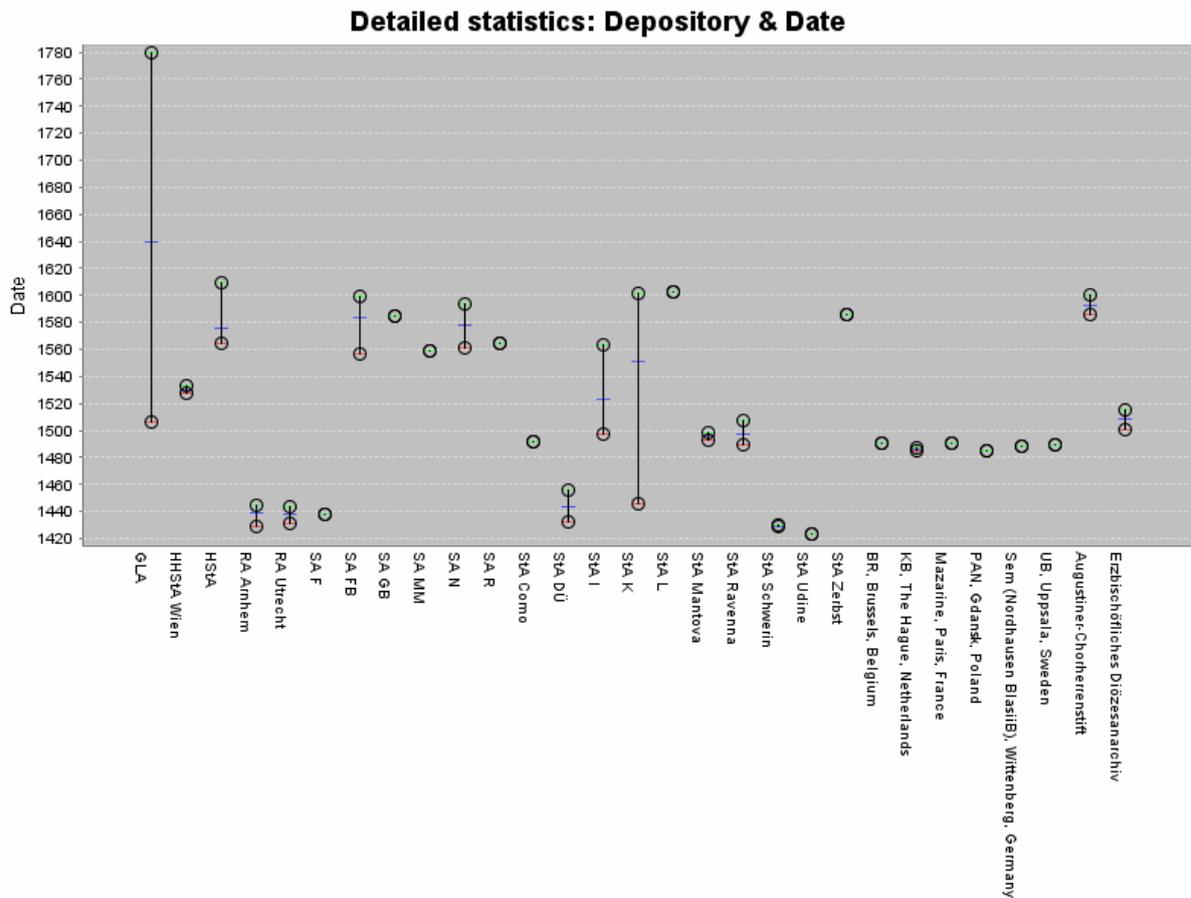


Figure 14: Detailed statistics: Depository & Date

11. Date & Height

The diagram for the “Date & Height” detailed statistics will be a “MinMax” chart similar to Figure 14.

12. Date & Distance

The diagram for these detailed statistics will be a “MinMax” chart analogous to Figure 14.

5 Handling of dates

Dates in the Bernstein databases either consist of a range with start and end year or can be treated as a single year value (if start year is equal to end year or only a start year exists).

For the detailed statistics with dates the following rules are applied: Single years are weighted with 1 and all years within a date range are weighted with $1/n$ (number of years in the date range).

We will demonstrate the above named rules with an example and assume the following dates were found:

1 x 1490	(weight = 1)
5 x 1491	(weight = 1)
1 x 1491 – 1494	(weight = $1/4 = 0.25$)
3 x 1492	(weight = 1)
6 x 1492 - 1493	(weight = $1/2 = 0.50$)

In this case the following “weights” for the years from 1490 until 1494 will be calculated:

1490:	1.00 (1 x 1.00)
1491:	5.25 (5 x 1.00 + 1 x 0.25)
1492:	6.25 (1 x 0.25 + 3 x 1.00 + 6 x 0.50)
1493:	3.25 (1 x 0.25 + 6 x 0.50)
1494:	0.25 (1 x 0.25)

List of figures

Figure 1: Summary statistics	6
Figure 2: Detailed statistics: Only Motif.....	8
Figure 3: Detailed statistics: Only Place of Use.....	9
Figure 4: Detailed statistics: Only Depository	10
Figure 5: Detailed statistics: Only Date	11
Figure 6: Detailed statistics: Only Height.....	12
Figure 7: Detailed statistics: Only Distance.....	13
Figure 8: Detailed statistics: Motif & Place of Use	14
Figure 9: Detailed statistics: Motif & Date	15
Figure 10: Detailed statistics: Motif & Height.....	16
Figure 11: Detailed statistics: Motif & Distance	17
Figure 12: Detailed statistics: Place & Height	18
Figure 13: Detailed statistics: Place & Distance	19
Figure 14: Detailed statistics: Depository & Date	20