

Documentation of the Online Indus Writing Database

Andreas Fuls, Berlin 2010

In most selection boxes you can choose one text. In some cases like **site(-s)** or **type(-s)** multiple selection/deselection is possible by holding down the STRG-key. The SHIFT-key selects a whole range of entries from the last selection to the point where the list is clicked.

The output boxes offers **HTML output** with click-boxes and **Text output** that are useful in copying the results to a Word document. Small pictures are useful in cases where small files are required. The high resolution of pictures is only necessary for publications.

Login

There are two levels of security.

First Login

The first login gives access to the user menu. The login is valid as long as the browser is open.

Second Login

The second login gives access to the database. You can login as a guest or with a user name and password, which are stored in cookies. Cookies will expire after 10 hours, but you can always login again (see main menu on the left side).

To register as a user contact the administrator. The number of data listed is greater for registered users as for a guest (limit of 10 data). If you login for the first time (after registration) choose any password, which will be stored into the database.

Information Window

Enter a **sign number** or **ICIT ID** in the navigation window to display detailed information in the information window.

Sign No.

Lists sign number, class, graph, type, number of varieties, frequency, function, ligatur, and comment. Click on frequency to see detailed information about the frequency of the sign. Click on function or ligatur to see and edit assigned functions, possible ligatures and their components, or to edit the comment.

ICIT ID

Shows an image (if available), lists archaeological and epigraphic information, classification, and displays the coded text with sign numbers. A new classification can be selected and saved (only editor and administrator).

Sign list

The sign list from Bryan Wells (2006) is the result of his PhD project and contains 676 signs. Each sign has a 3-digit code. The signs are grouped into different classes and types. To search through the sign list for a specific sign they are also grouped by their approximate graphic similarity. In most cases this will help you to find for example a sign that looks like a fish, a human or a bow (among other shapes). Of course, these graphic descriptions have nothing to do with the meaning of the signs.

Sign Functions

Some signs are assigned to a preliminary function as follows:

ITM	Initial Cluster Terminal Marker
LOG	Logogram
LON	Long Numeral
MIN	Minas
NUM	Numeral
PTM	Post Terminal Marker
SHN	Short Numeral
SSN	Short Stacked Numeral
SPN	Special Numeral
SYL	Syllable
TMK	Terminal Marker
XXX	Test function

Each code has exactly three characters, the same length as the 3-digit code of sign numbers. The function codes can be used together with the sign code during the [database search](#) or [matrix analysis](#).

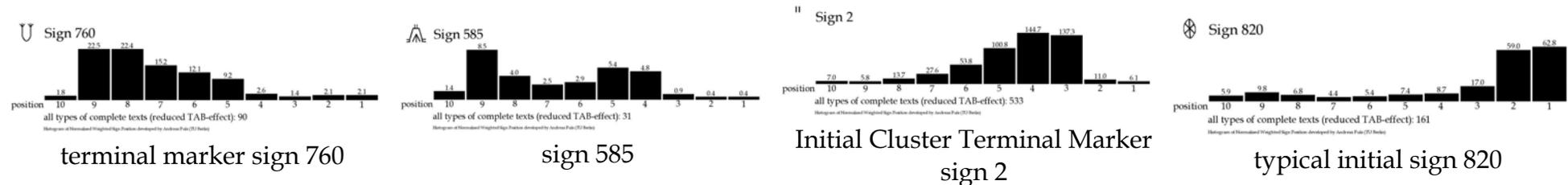
Sign Frequency

The frequency of one sign is listed for different parameters:

- Total frequency
- Frequency by text position
- Frequency by site
- Frequency by artefact type
- Frequency by time (only for Harappa, Mohenjo-daro, and Lothal)

It is possible to remove identical texts, thereby eliminating the so-called "TAB-effect". Another option is the verbose output, which lists the sign's frequency depending on the text position before one of the three Initial Cluster Terminal Marker (ICTM) , , and . A special feature is the output of histograms of the positional frequency for different text classes.

Sign position histograms



Sign Grouping

The grouping function is searching for signs, which have the same or similar sign neighbors, either before (preceeding) or after (following) the signs to group. The similarity is measured by the weighted euklidian distances between normalised sign pair frequency curves. This means, that their paradigmatic behaviour is very similar.

The method is performed in two steps automatically. In a first step the sign pair with the smallest normalised distance and not used before create the seed of a new group. All signs with a similar pair frequency and thereby a normalised distance smaller than the tolerance are inserted into the group. This process is performed until the maximum number of groups is reached or the smallest distance is greater than the tolerance. For each group the centroid with the mean values for each normalized sign pair frequency is calculated. At the end for each group the sign neighbors are listed in decending order of the number of corresponding sign pairs (N,

with $N > 1$, given in brackets). To avoid overlapping of sign groups sharing identical signs click on the check box **unique sign usage (disjunctive clusters)**.

The analysed dataset can be reduced by choosing specific sites, object types, a text class, only complete texts, and by reducing the TAB-effect. The remaining sign list can be reduced by the minimum number of occurrences for each sign. The maximum number of groups can be selected and a tolerance. The smaller the tolerance the smaller the number of signs in each group. The following example analyses long patterned texts from Mohenjo-daro reduced by TAB-effect with a minimum of 4 for each sign and a tolerance of 0.1 grouped by the following signs with disjunctive clusters.

Grouping of signs by their next sign frequencies (feature distance to centroid)

GROUP 1					
	100	690	760	585	923
	(0.025)	(0.025)	(0.025)	(0.097)	(0.097)
GROUP 2					
	705	706			
	(0.016)	(0.012)			
GROUP 3					
	817	861	820		
	(0.020)	(0.024)	(0.046)		
GROUP 4					
	32	415			
	(0.065)	(0.084)			
GROUP 5					
	388	390			
	(0.115)	(0.094)			

In this example each sign was used only once, thereby the clusters are disjunctive. It is possible to use signs more than once until they are used as a seed for a new cluster. To do so, uncheck the check box **unique sign usage (disjunctive clusters)** and choose a big tolerance. In the output groups can overlap and completely overlapping groups are marked as a subgroup of a previous group. It is also possible to get the output of the sign position histograms for each group. This allows to control the positional behaviour of the signs in each group.

Search

The ICIT is full searchable for any sign combination. All sign numbers have three digits (e.g. 017) and are separated by -. The Initial sign (0 in the example 231+) and the Terminal sign (1 in the example +407) are marked by +.

Place holders for one character is _ (underline) and % (percent) for any number of characters. This means, for example, that the search string %-585-___-231+ would look for any text starting from right with sign 231, the next sign to the left is unknown (three underlines to substitute any 3-digit code), followed by sign 585, and an open end of the text. In this example you will get three results from Mohenjo-daro as follows:

```
+407-032-520-100-585-017-231+  
+407-002-520-095-585-017-231+  
+407-032-520-095-585-017-231+
```

Other placeholders are [sign functions](#) encoded with 3 characters, e.g. NUM (Number), ITM (Initial Cluster Terminal Marker), and TMK (Terminal Marker). They can be used instead of a three-digit sign number, e.g. %+TMK-NUM-___-___-ITM%817% get three results from Mohenjo-daro as follows:

```
+520-033-706-233-002-817-798-000+  
+000-000-233-235-002-817+  
+000-000-000-798-002-817-032-320+
```

A text can have more than one textpart (separated in the sign code by +), e.g. the text code +144+700-033+ consists of two textparts with altogether three signs. It is possible to search for a specific number of signs and textparts by the use of =, <=, >=, <> (the option 'any' disables this search parameter). Sign 000 indicates one eroded sign, while ++ indicates an eroded textpart of unknown length.

The [sites](#) are grouped by [regions](#). It is possible to select one or more sites/regions. You should select either regions or sites, not both, since region and site have to match.

Output As

For the output it is possible to select short or long lists with a different number of information for each text, and to choose between HTML or text output, structured or unstructured text output.

A check box allows you to select the additional output of the **frequency between sign neighbors**. The frequency will appear between the images as small numbers. They are useful to determine possible word borders, since neighboring sign combinations with a high frequency are more likely words as less frequent sign combinations. If the sign combination not only appears with other signs but also exists alone as a solo pair, the frequency is highlighted in bold.

HTML

HTML output displays the ICIT ID as a button. Click on the button to display detailed information including images of the artefact in the [information window](#).

The option **HTML long** lists almost all available information for each text. The option **HTML short** minimizes the list to important informations. The option **HTML structured** lists the signs in different fields according to the search code entered into the input field Text (R/L). The option is useful to search for left and right adjacent signs. The **Order By** options field 1, field 2, etc. are only interpreted for structured output.

Example: Search for %033-705%803% and order by field 3, output as HTML structured

ID	site	type	excavation idno	field 1	field 2	field 3	field 4	field 5
1202	Mohenjo-daro	TAB:B	E232 CXVI:6	 +520-	 033-705	-	 803	 -853+
1254	Mohenjo-daro	SEAL:S	DK1104	 +740-760-	 033-705	-	 803	 -002-920-317+
815	Mohenjo-daro	SEAL:S	SD2445 105	 +520-	 033-705	 -220-	 803	 -055-002-820+
814	Mohenjo-daro	SEAL:S	VS1779 104	 +740-773-	 033-705	 -233-	 803	 -002-861+

1470	Mohenjo-daro	SEAL:S	DK-i 1066	 +520-070-	 033-705	 -233-235-	 803	+
404	Mohenjo-daro	SEAL:S	DK6357 446	 +740-585-017-	 033-705	 -233-798-	 803	 -002-861-603+
856	Mohenjo-daro	SEAL:S	DK221 147	 +520-	 033-705	 -236-	 803	+
385	Mohenjo-daro	SEAL:S	DK5821 426	 +740-176-	 033-705	 -240-235-	 803	+

Text

Text output is useful to copy the table to a Word-file (without any button). The option **Text** lists the most important informations and small sign images. The option **Text, highres** uses high resolution sign images useful for publications. The option **Textcode only** shows no sign images only the sign code.

Region list

All [regions](#) are listed in alphabetic order together with all sites.

Site list

All [sites](#) are listed together with their region, approximate settlement area, estimated begin and end of settlement, the total number of texts, and the total number of signs on the texts found in each site.

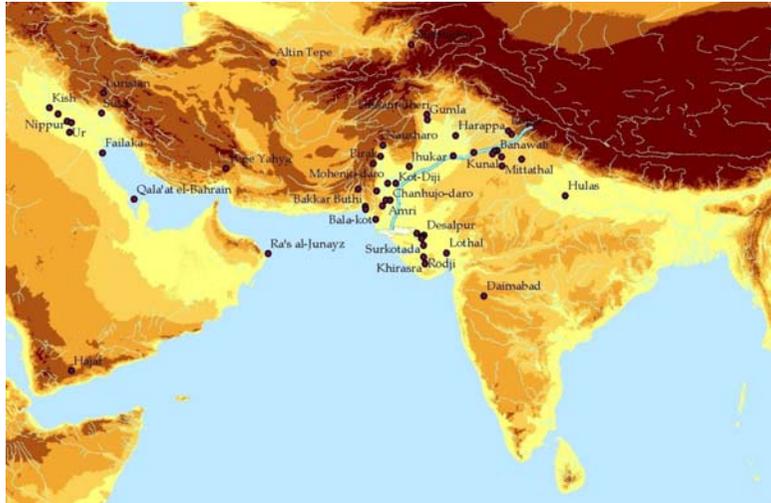
Time

It is also possible to search for texts of a specific time frame. The name of the [time periods](#) are defined in the menu [Time Line](#). There it is possible to assign a specific period and phase of a location (site, area) to a time period. This allows to search through one or many time periods. Please note, that not all texts are assigned to a time period (if a period or phase is not known or to inaccurate), therefore this reduces the set of texts.

Map of the Indus Culture

You can choose between two maps, an overview map and a map of the Indus valley. Both maps can be used to show the frequency of texts and their spatial distribution. You can either show the frequency of all texts or a subset of texts selected by specific parameters.

Overview Map



Map of the Indus valley



In the navigation frame (left) you can choose all parameters. Click on MAP to start or refresh the mapping. If the map shown is always the same (with different parameters) you should reload the frame in your browser manually (*or* disable e.g. in Firefox the extension tool "JAVA quick start").

Mapping the Frequency

There are many possibilities to calculate and display the frequency of texts in the map (for the [frequency of a single sign](#) go to menu [frequency](#)). In all cases a circle is drawn around each site with a frequency greater than zero, but the size of the circle depends on the following parameters:

Calculation of Frequency

There are generally two methods to calculate the frequency, absolute and relative frequency. The relative frequency depends on the total number to which it is related. It is either the number of texts per site, or the number of all texts which fulfill the search criteria.

absolute: absolute frequency of texts.

% texts / site: relative frequency of texts in relation to the total number of texts of the site.

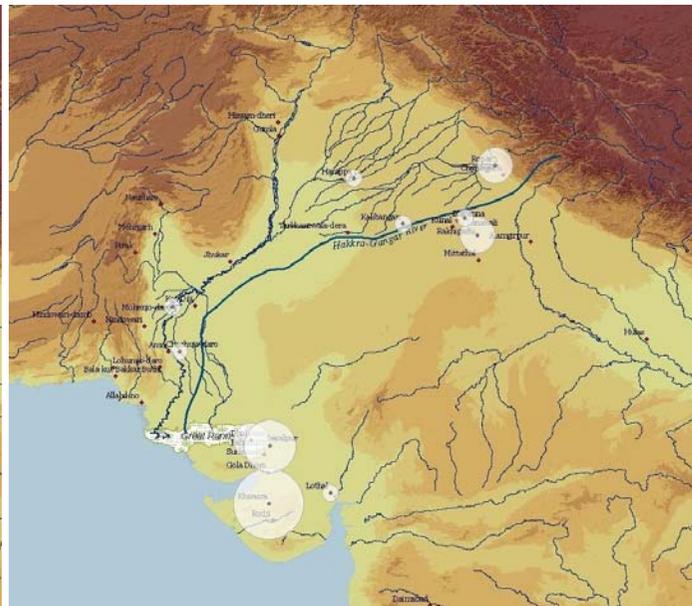
% of search result: relative frequency of texts in relation to the total number of texts that fulfill the search criteria.

Because of the irregular geographic distribution of texts (depending on archaeological projects) each method of calculating the frequency is problematic and has its own advantages and disadvantages. Since most artifacts with texts were found in Mohenjo-daro and Harappa, the absolute frequency almost always emphasize these sites and sites with a small sample of texts become almost invisible. The same is true for calculating the relative frequency in relation to the search result. In contrast, the relative frequency per site emphasizes sites with small text sample, since each text which fulfils the criteria has a very high weight in a small sample (sites with only one text get 100%).

Seal:R distribution (grouped into 4 classes)



absolute frequency of texts



relative frequency of texts per site

Display Function

It is important to understand how different frequencies are transformed into circles of different sizes, so that the map can be read properly. The value for each frequency is either absolute or relative, depending on the last parameter.

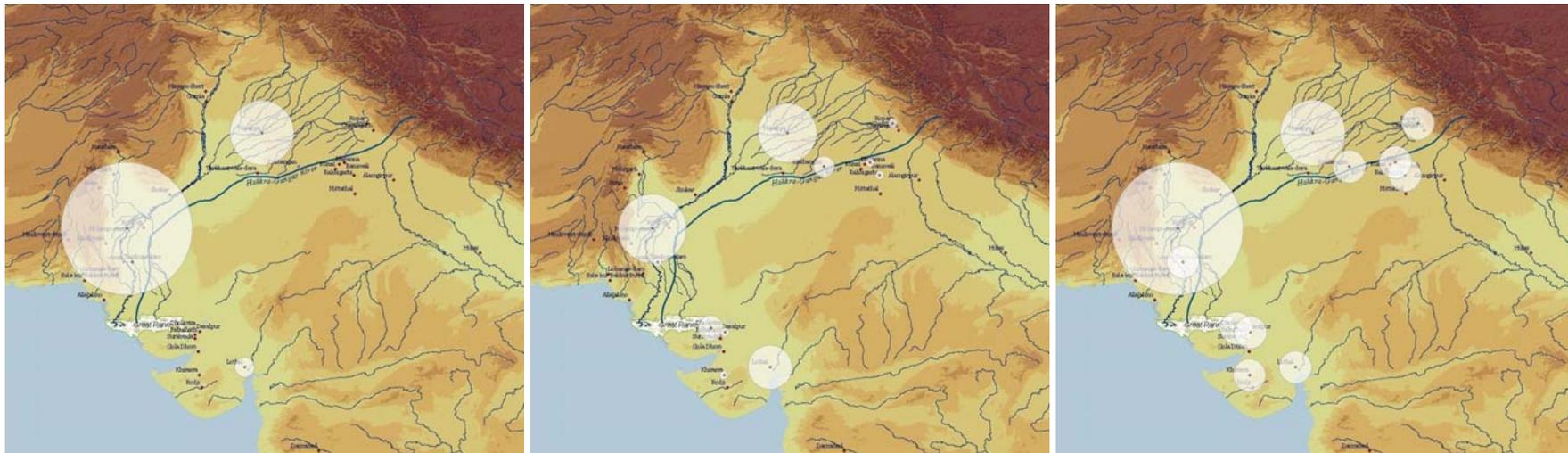
linear: circle size is proportional to the value of frequency.

logarithmically: circle size is logarithmically related to the value of frequency.

equidistant classes: All values of frequencies are grouped into 3 to 6 equidistant classes. This means that the maximum frequency is drawn with the maximum circle size, while the circle size of the lowest frequency class (class 1) is drawn with the smallest circle (radius = max size / number of classes). The radius for class 2 is twice as that of class 1, for class 3 triply of class 1, and so on.

The linear display function is easy to understand, but it has the disadvantage that small frequencies are almost invisible. The logarithmic scaling of the circle size emphasize small frequencies and reduces the circle size of high frequencies. Similar to logarithmic scaling is the equidistant classification, which also allows to display frequencies of a wide range, but it smooth away small differences.

Seal:R distribution (absolute frequency)



linear

logarithmically

grouped into 4 equidistant classes

Maximum Circle Size

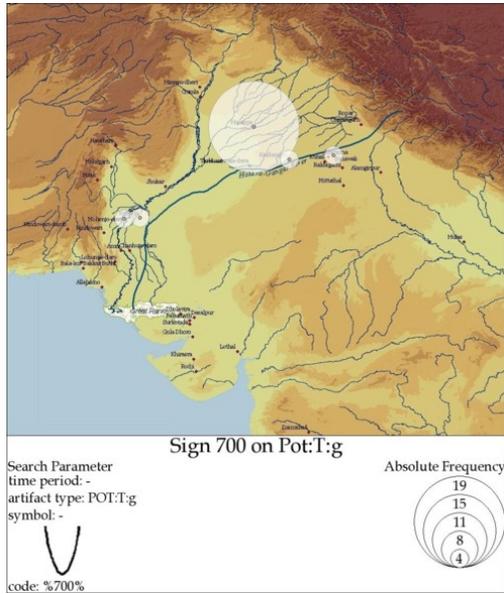
The parameter **max size** determines the maximum radius of a circle for the maximum value of frequency (absolute or relative). The circle size of smaller values is interpolated according to the display function.

Format of the Map

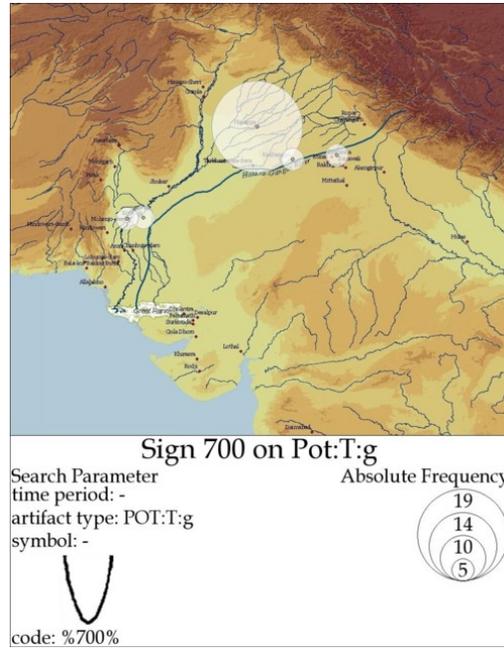
At the present only raster images are available. It is planned to include SVG in the future.

User defined Mapping options

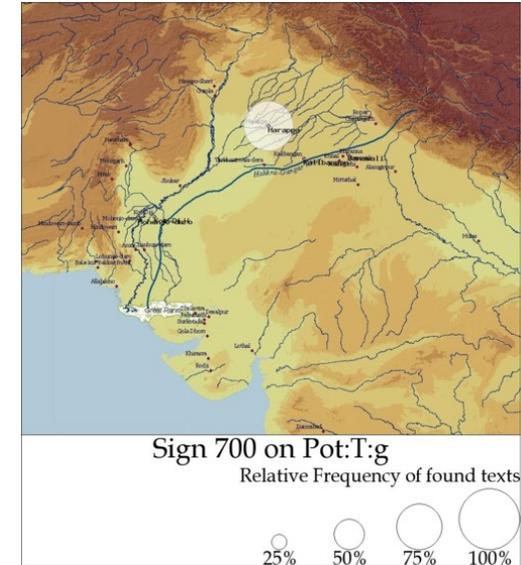
The map can be extended by different drawing options.



medium text size, absolute frequency,
grouped into 5 equidistant classes



big text size, absolute frequency, grouped
into 4 equidistant classes



big text size, relative frequency of found texts,
only legend of linear circle sizes

Title

The user can enter a title of up to 50 characters. The title is drawn below the map.

Search Parameters

The selected search parameters are listed on the left side below the title, if the check box "criteria" is checked. If a search includes sign code, the code and the images of the signs will be drawn.

Legend

To draw the legend, the check box "legend" must be checked. The legend gives the chosen calculation method of the frequency (absolute or relative by texts or by site), a list of circles with different diameter according to the [Maximum Circle Size](#) and [display function](#), and the range of values. The number of circles in the legend is equal to the number of classes or 4.

Site Name

The check box decides if the name of sites with found texts are drawn on the map or not.

Region Name

Not yet installed.

Text Size

The text size is a factor for the size of site names, the text size of the title and the legend text.

Structure analysis

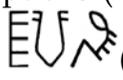
The structure analysis of texts is similar to the search tool. It allows to search for texts with signs in a specific text position before or after one or more Terminal Marker(s) (TMK) and before or after one or more Initial Cluster Terminal Marker (ITM). Additionally, the exact position of the ITM can be chosen.

For example, sign 920 after ITM 060 in position 1 to 4 and sign 220 afterwards and before the TMK 520 or 740 results into 4 inscriptions:

+520-220-415-060-920+
+740-690-435-255-240-220-415-1006-742-060-920+
+520-220-235-060-920+
+520-220-003-741-060-920+

This tool has a beta-status and need to be checked for correct results. The use of sign function codes does not give correct results. Therefore, the user should compare it with results of the [search tool](#).

Sign cluster analysis

The sign cluster analysis searches through the corpus of Indus texts for two or more signs with a minimum number of occurrences. All sign combinations are counted how often they occur in the inscriptions (For example, in the text "ABCABGAC" the sequence "AB" occurs twice). It is a tool to discover repeating sign sequences like  (400-740-176), which occurs 36 times on "TAB:B", "TAB:C", and "TAB:I" from Harappa. Having a list of sign clusters it is possible to check each sequence of signs in the search menu and to analyze them as potential sign sequences representing words.

The cluster analysis is very time consuming and the search for more than 3-sign clusters takes not only a long time but also often results in an empty list. The search of 3-sign clusters in 3169 different texts takes about 45 minutes. It is recommended to start with clusters of two signs and extending the analysis to longer sign cluster afterwards, by either using this tool or manually. It is also recommended to reduce the number of analyzed texts by the selection of one site or artefact type, thereby speeding up the cluster analysis.

For example, the search for two sign clusters on POT:T:g artefacts searches through 322 texts and takes about 15 seconds. With a minimum of 4 occurrences the resulting list of sign combinations are as follows:

frequency	text code	sign cluster	occurrence in ICIT
4	001-410		1696, 3639, 3728, 3729
6	700-700		2771, 2994, 2995, 3192, 3220, 3223
4	002-861		1692, 3127, 3630, 3684

Reducing identical texts from the same site to one leads to 226 different texts and only signs 700-700 and 002-861 appear both (at least) 4 times. It is obviously, that in this case a search for more than a 2-sign cluster results in an empty list.

Analysis

This is a collection of special analysis methods.

Logograms

Search for texts with single signs (+__+), which are candidates of logograms. The resulting list gives for each found sign the frequency and lists all related artefact types.

Numbers

Search for texts with numbers in any text position. The result is identical to the search for %NUM%.

Positional Numbers

Search for texts with two adjacent numbers in any text position. The result is identical to the search for %NUM-NUM%.

Number combinations

Lists a matrix of frequencies for number combinations (%NUM-NUM%). It is a special tool of the more general [matrix](#) tool.

Calculation

Search for texts with numbers in two separated text parts. The result is identical to the search for %NUM%+%NUM%.

Matrix

Here it is possible to count the frequency of two-sign combinations and get an overview as a matrix. Sign 1 (left) and sign 2 (right) can be separated by

% any number of signs or text code

- no sign, which means they are adjacent to each other (default option)
- + a text end/beginning marker, which means they are in two different text parts

For example, the search for TMK-NUM count the frequency of all Terminal Marker (TMK) right adjacent to all numbers (NUM). The following example shows all combinations of Terminal Markers (TMK-TMK):

-		right sign					
left sign	no.	90	226	400	520	527	740
		𑀓	𑀔	𑀕	𑀖	𑀗	𑀘
𑀓	90	-	-	1	1	-	115
𑀔	226	-	-	-	-	-	-
𑀕	400	12	3	-	15	3	216
𑀖	520	-	-	8	-	-	-
𑀗	527	-	-	-	-	-	-
𑀘	740	1	-	14	-	-	1

The background color is only set if the code for Sign 1 and for Sign 2 are identical (e.g. TMK-TMK, or NUM-NUM). If you search for different sign combinations like NUM-ITM or 700-NUM no background color is used.

-		right sign																															
left sign	no.	1	2	3	4	5	6	7		13	14	15	16	17	18	19	20	31	32	33	34	35	36	37	39	55	65	415	416	705	900	909	
		'	"																									𑀓	𑀔	𑀕	𑀖)	𑀘
𑀓	700	-	-	2	8	-	-	-	-	-	1	-	-	-	-	-	-	3	89	136	105	-	2	-	-	-	-	-	-	-	-	-	

References

Wells, Bryan: An Introduction to Indus Writing. MA Thesis, Early Site Research Foundation, Independence, 1999.

Wells, Bryan: Epigraphic Approaches To Indus Writing. Dissertation, Harvard University, 2006.

Citation of this document and the database

Wells, Bryan, and Fuls, Andreas: Online Indus Writing Database. Berlin 2010. +URL (+date of last access).

Fuls, Andreas: Documentation of the Online Indus Writing Database. Berlin 2010. +URL (+date of last access).

