

decodeunicode.org – the open science database

Prof. Johannes Bergerhausen

University of Applied Sciences Mainz, Germany

Keynote Address, IUC27, Berlin, Germany

Thank you very much for inviting me to present the project decodeunicode in the keynote to today's conference – it is an honour for me to be here. decodeunicode is a project initiated by the Department of Design at the University of Applied Sciences in the German city of Mainz and, as such, is supported by the German Federal Ministry of Education and Research.

In the first part of my talk I would like to give a brief overview of the history of character encoding from telex to Unicode – seen through the eyes of a typographer ... followed, in the second part, by an introduction to the project decodeunicode.

1.1 The BMP Poster

Allow me to begin with a practical experience we made during our work on the project. One fine day, in our mailbox we found a simple rtf document sent by the programmer of our database, Wenzel S. Spingler. At 512 K, the document was rather small in size, but it contained an idea both simple and beautiful.

With the help of an AppleScript, Wenzel had typeset in one document all 65,536 characters of the Basic Multilingual Plane (BMP) – one after the other, code point by code point. This, indeed, looked interesting enough for us to decide and make a poster of it which would retain a maximum of the rather special look the original raw text data had.

One of our team members spent well over 70 working hours to assign the different fonts to the code points. Working in InDesign CS, he decided to give each font its own colour so as to not lose himself completely – which was no easy task, considering he finished by using 17 different fonts.

What you see here is a detail of the resulting poster, with the imprint designed like a header. The colour underlying the glyphs resembles the marking of a text and communicates the fact that all these characters are available on today's computers.

To ensure legibility of the characters, we agreed upon a font size of 15 point. I must admit that the ensuing needed size of the poster took us by surprise: After discarding German standard poster formats like DIN A0 or 70 by 100 cm as too small, we finally managed to fit it all into the so-called CityLight format used for advertising in public places which is 1 meter 20 by 1 meter 80.

This picture shows the printing plate. We produced 300 copies, with proceeds benefiting the project.

To me, the poster is a good illustration of the tasks typographers in the 21st century are facing. Probably the nicest compliment it was paid came from an Italian typographer who called the poster an alternative world map.

With one of the main tasks for us communication designers being to communicate things as clearly and easily as possible, in our case, the result speaks for itself. The beauty is in the Code. All we had to do was to render visible what was already there.

1.2 Character Encoding from a Typographic Point of View

One thing has been intriguing me ever since, in 1997/1998, I carried out a typographic research project¹ on the ASCII Code for the French Centre National des Arts Plastiques in Paris. How is it done that when I press a key on my keyboard, that same character will appear on some screen, ink jet printer or CTP printing plate at the other end of the world? There is, as all present here will be aware, no easy answer to this. My professional background is not a technological one and I am unable to program but I am interested to know how technology influences typography.

One of the first systems for the transmission of written characters across a distance was the TELEprinter EXchange or TELEX machine, shown in this picture as a model from the 1920's – an important invention which, in principle, is in use until today in spite of great typographical limitations. With the first international telex codes allowing for no more than 5 bit, i.e. 32 code positions, there was, at the beginning, no room even for a distinction between upper- and lowercase letters.

This photography may look like a filmstill from a James Bond movie but is, in fact, a computer the United States Ministry of Defence used at the beginning of the 1960's. Just like the telex, the computer adopted the keyboard of the typewriter as input tool – with the exact QWERTY keyboard layout the inventor of the typewriter, Christopher Sholes, had filed as patent in 1867. Certainly a technological milestone, this was none typographically.

There is one group quite apart from the telecommunication and computer professionals who, like it or not, produce typography, namely the international army of secretaries and typists, in a way direct descendants of the court scribes. In 1961, the IBM typeball caused a sensation. This electric typewriting machine featured 88 characters and an option to switch fonts in mid-text by exchanging the typeball. While the typeball, thanks to its even touch, produced very clear letters, the monospace setting of all characters lead to a generally unbalanced look – certainly nothing a hot-metal typesetter would have considered as typography.

The IBM Magnetic Type Selectric Typewriter introduced in 1964 basically marked the invention of modern word processing. At that time, the idea of word processing analogous to data processing was completely novel. With the help of a two-tape machine, the Selectric was able to store 24 pages of addresses and text costing the equivalent of about 20,000 euros. It thus marked the birth of the serial letter and became a huge success.

The machine used a novel character code called ASCII or, in its original version, "US American Code for Information Interchange"² developed for the US American Standard Institute ANSI. At the beginning of the 1960's, IBM employee Bob Bemer headed a group of programmers working on making – and here I quote the Reader's Digest of June 1961 – "machines (...) talk with machines – languages that will facilitate the exchange of information by radio, microwave or telephone wire between computers at widely separated centres"³. In the year 1961, this was pure science fiction. Developing a common US American standard, Bob Bemer's team analysed over 60 different codes existing at the time.

Little did the team of engineers guess they were working on a future world standard when their first version of 1963 didn't provide for lowercase letters as the developers saw no need for them. Today, I like to think that the ASCII Code might well be one of the most successful codes after the DNA. Preserved over time in an almost evolutionary way, it has survived until today in the first block of the Unicode standard.

The ASCII Code chart bears not just a visual resemblance to Russian Dmitry Mendeleev's 1869 periodical system of chemical elements. Mendeleev had the inspiration to leave room in his chart for elements which had yet to be discovered!

This photography exemplifies typography on a computer at the dawn of the ASCII era – a rough and primitive screen with monospace letters. In the days of the typographic stone age, the subtleties of micro-typography didn't stand a chance.

In view of this, it comes as no surprise that typographers and typesetters of the time tended to look down on the computer condescendingly. Bear in mind that hot-metal setting, almost unchanged since the days of Gutenberg, had indeed set rather high standards. In hot-metal setting, the range of special characters available by far exceeded that of the computer. Trained typesetters would find the characters they needed almost without looking. In letter cases each character basically has already been assigned a fixed position on an x- and a y-axis.

But the computer industry of those days had other things on its mind than the subtleties of typography. Two great inventions would not have been possible without a common character set like ASCII. This picture illustrates the simplicity of a grand idea: Four universities – UCLA, UCSB, SRI and Utah – were the first nodes of the Arpa Network in 1969, which later was to become the Internet.

Ray Tomlinson, too, drew from ASCII when, in 1972, on a TeleType 33 he invented the e-mail. Looking for an immediately recognizable separator between the name and domain of an e-mail address, on his TeleType keyboard he came across the "commercial at" character. This ligature had been part of the ASCII Code since 1963, while its first evidence dates as far back as late 19th century America.

The so-called desktop revolution of the early 1980's isn't so much to be ascribed mainly to the Apple Macintosh as to the fact that Apple used a new page description language called PostScript which had been developed by a small start-up business called Adobe. When, in 1986, font producer Linotype followed the example of the Apple LaserWriter by also using PostScript fonts in its image-setter Linotronic 300, typography had finally and truly made it into the digital age.

The 8-bit codes of the 1980's and WYSIWYG computers lead to a big bang in type design. Almost overnight, graphic designers and amateurs alike were in a position to design characters themselves. To an increasing number of people, the secret lore of type design became a focus of interest.

There is, however, another aspect which seems to be of even greater importance. With the advent of the 8-bit codes, the completely different typographic output media of typewriter, computer printer and PrePress HighEnd imagesetter merged into one industrial branch. And by way of a kind of democratisation of characters, secretaries, scientists, CEOs, designers and amateurs of typography were now all using practically identical character codes.

But RankXerox was to take things yet another step further. As early as in 1981, the Xerox Star featured almost everything a modern personal computer does: Graphical User Interface, desktop metaphor, folder structure, icons, WYSIWYG and mouse. On top of that, the Star was the first computer worldwide with a 16-bit code designed to typeset and print scripts like Cyrillic, Greek or even Japanese characters. Sadly, the Xerox Star didn't have much commercial success.

Still, the idea of creating one common international character set remained a revolutionary vision. In 1984, Xerox developer Joseph D. Becker wrote: "The fascinating variety of human writing systems must be brought to co-existence in the computer. At first sight, this hardly seems possible. Arabic, for instance, flows in flourishes from right to left. Thai and other ancient Indic scripts have characters

wrapping themselves around their neighbours, thus leaving the phonetic order. [...] Nevertheless, the ultimate goal must be multi-lingual word processing.”⁴

Before turning to Unicode, I’d like to illustrate with a chart the development of coded character sets in the 20th century. The number of encodable characters has risen exponentially. In this area, mankind has indeed made great progress over the last few years. While the early output equipment of the first available systems, namely telex, 6-bit codes and ASCII, didn’t have too much to offer in the way of high quality typography, today’s computer world has more characters at its disposal than Gutenberg would ever have dared dream of.

1.3 Unicode from a Typographic Point of View

When, after several years’ work, Joseph D. Becker and other members of Xerox and Apple in 1990 published Unicode 1.0, typography was about to enter a new era. Today, the fact that Unicode could in such a short time evolve from Utopian vision to world standard, seems as much a product of hard labour and good fortune as a crucial step forward in cultural history.

“We shape tools and they in turn shape us.”⁵ As it happens, Marshall McLuhan coined his famous phrase in the same year the ASCII code was developed – 1963. Typography has always been largely influenced by technology. Any achievement in this area has led to new designs. “The medium is the message”⁶, Marshall McLuhan might be quoted in a somewhat overstated way. It is my belief that apart from marking an enormous technological advancement, the Unicode standard before anything else represents cultural progress. With mankind working on, in a layman’s terms, a digital inventory of all characters of all cultures, the impact of Unicode on typography is only just beginning to show.

Characters are the “genes of communication”, and their common technological basis brings the world a little closer together. With the Unicode standard, typography has entered the age of globalisation. Until now, a type designer in the Western world used to design about 200 glyphs per font and wouldn’t even start thinking about a Cyrillic or Greek version unless his design had become highly successful. Nowadays, things are becoming a little more complex. Type designers will increasingly have to think internationally and look beyond the boundaries of their own culture. And since no single type designer can possibly know all the characteristics and cultural background of all writing systems, international teamwork is called for.

The line of code from UnicodeData.txt encodes the properties of Latin Capital Letter A. To me, this line of text ideally illustrates the difference between character and glyph. The character is entirely abstract and should not even be illustrated by a glyph. In a way, the character is an expression of a Platonist idea because it is devoid of all shape.

Glyphs, on the other hand, are the formal expressions of the one idea “Latin Capital Letter A” and, in principle, infinite in number. They are the type designer’s daily bread.

In the age of Unicode, the LastResort glyphs developed for Apple by script expert Michael Everson are a great help. Thanks to them, the block can be determined even when a font is missing. The system fonts provided today by ordinary operating systems cover a large number of Unicode characters. Nevertheless, some developing work remains to be done. The natural common goal should be for every computer on the market to provide as system font the entire number of characters listed in the latest version of the Unicode standard. To achieve this goal, producers of operating systems ought to collaborate with OpenSource projects or academic institutes to jointly develop any missing glyphs.

But the LastResort glyphs illustrate yet another aspect. Michael Everson was in fact as far-sighted as the Russian inventor of the periodical system of elements when he developed LastResort glyphs for writing systems that are as yet waiting to be encoded. The darker LastResorts in this picture represent writing systems listed in Unicode 4.0.

1.4 Cultural Background of a World Standard

Typography has always been political. Take the following example from modern history: This hand-coloured photographic document shows Turkish statesman Mustafa Kemal, known as Atatürk, who, in 1928, by decree made an entire population use the Latin alphabet instead of the Arabic characters. Less than a century later, in 2005, we witness the first round of negotiations on membership of Turkey in the European Union.

The choice of the United Nations flag as icon for the international keyboard layout of Apple OS X to me seems particularly appropriate. In a way, the Unicode standard is a typographic UN General Assembly in which, consequently, every writing culture ought to be entitled to a seat. This is why our team supports the Script Encoding Initiative launched by Deborah Anderson at the University of Berkeley, California.

It is in the same vein that I consider the development of keyboard layouts for all existing writing cultures such as Pashto for example, an extremely important cultural task. The access of cultures to modern communication technologies should not be dependant on the economic interests or, more precisely, the lack of such, on the part of the IT community.

It seems quite obvious to me that one day, all modern operating systems ought to be translated into every living language of the world. This may sound little more than a utopian dream today, but the situation may well change in the future. In Cambodia, for instance, the Khmer OS Initiative is seeking to have future operating systems localized in Khmer. Can anyone think of a nobler development project?

A look at the Japanese keyboard illustrates typographic globalisation today: To be found here are Katakana characters, Arabic-Indic numbers and Latin letters – a melting-pot! Script, one of the most formidable cultural achievements of mankind, has been developing over millennia in many countries and by many a people. This very development, to my opinion, will not, as some people see the writing on the wall, lead to all mankind using but the Latin alphabet. The contrary is the case – as an extremely robust and democratic instrument, Unicode will ensure a peaceful co-existence of the diversity of written cultures on the keyboard.

Ever since the days of hot-metal setting, a character accidentally appearing in a text and standing out in different style or weight than the rest has been called, as we say in Germany “Zwiebelfisch” (“onion fish”, “literal mistake”). Readers with a perceptive eye will instantly experience irritation at the sight of such inaccuracy.

This phenomenon lives on in many modern publications whenever designers or type-setters can't find the corresponding Non-Latin script. I would like to call this kind of mistake “international zwie-belfisch”. Luckily, an increasing number of well-developed fonts cover more than one writing system.

From a typographic point of view, the reference glyph ought by all means to consist of a shape as monolinear as possible, thus focussing on the essence of the character.

All the people who, during the past fifteen years, have worked for the advancement of the Unicode standard – and many of them are present today – have done a tremendous job. I have no doubt that this success story will continue in future versions. The world standard has been firmly established and, in a way, already turned into a cultural world heritage. We are constantly becoming more aware of the fact that it is one world we live in. The tools we shape are indeed shaping us.

Still, a lot remains to be done. The communications media of the Unicode Consortium so far address an exclusively expert audience. The publication “The Unicode Standard 4.0” is a wonderful volume for programmers and encoders, but linguists, type designers, let alone ordinary computer users will find it hard to handle. In fact, only a relatively small part of the non-professional computer-using community knows anything at all about the Unicode world standard. Type and graphic designers, for their part, are becoming more aware of Unicode as many font labels are gradually switching to OpenType.

2.1 decodeunicode – Every Character Tells a Story

Characters represent the essence of a culture, literally conveying history in their shapes. Missing, however, is information about the names, history, different meanings and correct typographic use of thousands and thousands of characters. This I would like to illustrate with a few examples.

A large number of graphic designers use “Logical Not” as a bullet point for the simple reason that, thanks to ISO-Latin 1, it turns up in nearly every font. I know a few type designers who have designed it many times without knowing what it actually means. Is this not rather absurd? Any PC user in the world today has thousands more characters at his or her disposal than Gutenberg did, without knowing hardly anything at all about their existence or meaning.

Who actually invented the full stop? Whoever it was, he or she was an outstanding designer. Is there a historic link to the Japanese full stop? Access to and comparability of the different writing systems will lead us to new questions.

The currency symbol, known by some as “sputnik”, and has a very curious story to tell. When, at the end of the 1960’s, the ASCII code was to be made international ISO standard, representatives from a number of countries refused to consent to the US-American dollar character being included in ISO 646. The Cold War, apparently, wouldn’t stop at typography. To resolve the problem, it was decided to design one universal character for the term “currency”. Rarely ever used to date, “currency” nevertheless still figures in the special character repertoires of many a modern mobile phone. The symbol represents a coin reflecting the sunlight.

What you see in this picture is not a Japanese character. With each of the three Japanese writing systems having a Yen character of its own, this may well be a Latin letter “Y” with two horizontal strokes added on the analogy of some “dollar” or “pound Sterling” glyphs. Who invented this character which first appeared around 1910?

The exclamation mark made its first appearance at the beginning of the 16th century and is today firmly established and known all over the world. Depending on the context it is used in – at the end of a sentence, in mathematics, in a comic strip balloon or on a road sign –, it carries various different meanings. Now is the time to collect them all in one interdisciplinary effort!

The two characters a and A bridge a period of over 800 years. The Romans adopted the upper case “A” from the Greeks. Logographically, it was derived from the early Semitic “aleph” for “cattle” and

when you turn it upside down, it still resembles a cow's head. The common lowercase "a" originates from the Caroline minuscule developed during the reign of Charlemagne.

The pictogram for airplane or airport will certainly be understood all over the world. Unicode rather in passing encoded a number of pictograms in its "Dingbats" and "Miscellaneous Symbols" blocks. The international symbol language of orientation systems ought to be systematized and represented in future versions of the Unicode standard.

The French "Guillemets" shown in this picture have been ascribed to printer and punch cutter Guillaume Le Bé and were first used in 1527 Paris. Of a round shape, they at first resembled a double bracket at x-height.

This final glyph is an Arabic ligature translating into "May Allah's peace and blessings be upon him". All these characters have a cultural background that might be of interest to a broader public audience. Or, in brief: Every character tells a story.

2.2 decodeunicode – The project's concept

What decodeunicode.org wants to do is to collect these stories and many more. It is my belief that in the years to come, access to the Unicode characters will increasingly become a normality for typographers, designers, scientists and laypersons alike. As a consequence, more and more questions will be asked as to their meaning. Therefore, the code must be decoded! Before giving a first impression of what the decodeunicode.org website looks like, allow me to outline the conceptual basics of our project.

Structure. Strictly based on Unicode blocks and hexadecimal values, the database offers a clear structure. In its first version, it reflects the complete Basic Multilingual Plane.

Basically Encyclopaedic. The Unicode standard is situated at the interface of a large number of disciplines all of which we want to call in for the project. It is our goal for the database to provide information to semiotics, typographers, linguists, communication scientists, computer scientists, mathematicians, historians, palaeographers as much as to an interested lay audience. If we succeed in doing just that within the next months and years, we shall be developing nothing more and nothing less than an encyclopaedia of Unicode characters.

Typographic Enlightenment. We want to show the typographically correct shape and use of every character or, in type designers' terms, the "design guidelines" defining the basic principles for the design of the ideal glyph. To many of today's type designers who design characters originating from another culture than their own, this may indeed be a very useful resource. Users and designers, too, may find this kind of information interesting. What ought a Polish ogonek to look like? What does this mathematical symbol mean? How should I use the various existing quotation marks?

Technical Background. The database shows the official Unicode properties of each character. These values may also be accessed via a search function.

Intuitive Surface, Scientific Depth. The clearly laid out surface presents complex correlations – as easily as possible. For each character, scientific information on various topics like history, spreading, meaning etc. can be collected. Should need arise, a discussion platform can be opened for each character.

Open Science. There isn't a project team in the world able to write 65,000 articles within any reas-

nable period of time. Even if we were to deduct from this number a couple of thousand Not Assigned and Private Use characters, we are acutely aware of the fact that Version 1.0 of our project can be nothing more than a beginning and will leave quite a bit of work to be done in the future. It is for this reason that we made decodeunicode an OpenScience project. Every Internet user from professional to amateur can insert, write, edit or add to an article or even upload images on every character onto the website. This may sound rather daring at first but has in fact been proven very successful in the OpenScience project of the free encyclopaedia at www.wikipedia.org⁷. By the beginning of 2005, the English version of wikipedia alone featured sound knowledge about the world in roughly 445,000 articles.

Free. True to the Open Science creed, it goes without saying that the enlightenment on the characters of the world is free of charge. Any published material is freely reproducible according to the Creative Common Licence⁸ prohibiting only purely commercial use.

English, German. Version 1.0 of decodeunicode is published in two languages, German and English. Support for any other languages is most welcome.

Sponsors welcome. Due to our extremely small-sized structure at the University of Applied Sciences in Mainz and the 18-month limit of sponsorship for our project we are forced to content ourselves with an initial version 1.0 of our database. We would most gladly welcome further sponsorship.

Should anyone like to learn more about the concept of our project, there are brochures available at the conference bookstore.

2.3 decodeunicode – The Website

To conclude my talk today, let me illustrate our project by showing you some screen shots of our decodeunicode.org website. Simple and intuitive access to the blocks of the BMP is provided by, what we are calling the BlockDock. This flash navigation depicting one glyph each as representative for the 105 blocks allows a playful discovery of unknown characters and cultures.

This navigation leads to an overview of any selected block. The pictures of the reference glyphs without exception are .gif documents, thus making sure even users lacking the corresponding font in their system can view the glyphs. The bottom screen area offers anyone the opportunity to edit information on the selected block. One thumbnail per character leads to yet another screen dealing exclusively with this particular character.

The glyph-page is the heart of our website. Every character tells a story. It was important to us to include a large image of the reference glyph. The arrows pointing to the left and right make for a comfortable navigation through the material on single characters. This clearly illustrates the basic principle underlying the design of our website. The top half of the screen is strictly black and white and concentrates on the reference glyph and further information. This is official Unicode information which may not be edited. In the coloured bottom area users have the opportunity to access and enter information themselves. This distinct structure helps tackle the very complex subject we are dealing with.

Each code position now allows for entries on the following subjects: Intro, History, Spreading, Meaning, Typography, Discussion and Gallery. Click any of these editing buttons and you will be taken to the front end of the database where you can comfortably enter pieces of text or upload images. Upon doing so, one is reminded that any piece of text or image entered must be

copyright-free, i.e. open to be used by anyone. The Creative Common License draws a line of distinction only at commercial use. Having entered a text or edited an existing one, click "send" at the bottom and within a few seconds your entry will be online. As the content management system stores all previous versions of a given text, you may also follow the editing history of an article.

It is our firm belief that the power of the glyphs alone is enough to attract the viewers' interest and to make them enjoy browsing through this encyclopaedia of characters. This is why we included the option to further enlarge the view of a glyph in order to view nothing but the reference glyphs. With the resolution of monitors improving steadily, our website is designed to satisfy owners of 23" or 30" monitors as well.

Any character may include a collection of images – anything from an exemplification of an ideal letter to everyday typography. This option allows for an interesting glimpse into unknown character cultures.

Images of a chosen character can be viewed in a gallery which includes the option for the author to comment and, once again, the note to enter exclusively copyright-free material.

A full text search option simplifies the search for any particular block, character or property. This screen, for instance, shows the list of findings on the word "Mongolian".

To conclude my presentation I would like to say that our small but beautiful team has worked hard over fifteen months on version 1.0 of our vision. As you can guess from the version number, it certainly is some way from being altogether perfect.

Still, I am proud to tell you today that last Monday, the first version of decodeunicode has gone online. We would greatly appreciate your participation, entries and constructive commentary!

Thank you very much.

Cologne, Germany, January 14, 2005

English translation by Nicola Fischer, Heidelberg, Germany

- 1 Johannes Bergerhausen: Questionner le clavier: Les Codages ASCII et ISO-Latin 1, Dossier, CNAP, Paris 1999
- 2 Charles E. Mackenzie: Coded Character Sets, History and Development, Reading, MA, USA, Addison-Wesley, 1980, ISBN 0-201-14460-3
- 3 Bob Bemer, Reader's Digest, June 1961, quoted at www.bobbemer.com, accessed January 14, 2005
- 4 Joseph D. Becker: Multilingual Word Processing, Scientific American No. 251, July 1984, page 96-107, re-translated into English from a German translation of unknown origin.
- 5 Marshall McLuhan, Understanding Media – The Extensions of Man, New York, USA, McGraw Hill, 1964
- 6 Marshall McLuhan, Understanding Media – The Extensions of Man, New York, USA, McGraw Hill, 1964
- 7 www.wikipedia.org, accessed January 14, 2005
- 8 Creative Common Licence, <http://creativecommons.org>, accessed January 14, 2005