The Rosetta Disk Story
(or How to make a Rosetta Disk and back up or civilization)

by Alexander Rose
Executive Director, The Long Now Foundation

Back in 01998 the idea was put forward that a modern Rosetta Disk might be made using micro-etching technology. The disk would have as many translations of the same text as possible, all etched into metal that could last for thousands of years.

The first person I remember discussing this was Brewster Kahle of the Internet Archive at the Time & Bits conference. The idea was discussed further by Doug Carlston, Stewart Brand and myself on a long car trip in the southwest after hearing that portions of the bible had been translated into at least 2100 languages.

The technology for micro-etching that Brewster had mentioned was adapted by Los Alamos Labs from gallium-ion beam micro-circuitry FIB machines to etching text instead of circuits. They specifically developed this in order to store data for the nuclear waste storage program that had a congressional mandate of 10,000 years. This technology had been licensed exclusively by a company called Norsam to make high density archival materials, in fact they called it *HD Rosetta*. We contacted them and found out that they could theoretically put around 300,000 pages on a 2.8" metal disk, with the caveat however that you would need a very specialized electron microscope to read it. They estimated that at 10x that scale, or 30,000 pages per disk, the content would be readable with optical microscopes in the 1000x range (17th century technology). This seemed like a good target, and after hiring project director Jim Mason, and securing funding from Charles Butcher for both a conference and a disk, Long Now set about collecting parallel texts.

Then came the long discussion about what to collect and who to get it from. We knew that we could not translate any piece of content into thousands of languages ourselves, and so we would have to choose something already translated. We had found through research that the only document translated into more than a few hundred languages was the Bible (1), but after talking to many archives, very few had more than 10-20 translations themselves. And because most graduate and PHd students tend to get published for increasing specialization in linguistics, the academic materials for broad survey work was extremely limited (this is an effect we later saw with the All Species project as well). There was one exception, SIL or The Summer Institute of Linguistics. They are one of the largest non-profits on the planet, and chances are that you have never heard of them. They grew literally out of a summer camp to train young missionaries how to translate Bible materials into undocumented languages. This activity of course is fraught with all kinds of anthropologically sticky issues. We quickly learned however, that if we were to do this project, that SIL

would have to be involved.

SIL is not legally a religious organization, but is an academic linguistic institute that trains and supports missionary activities. Historically they were the only organization to produce broad survey data on all known languages. They publish their basic information on each language of the planet (around 7000) in an amazing tome called The Ethnologue (now in its 15th edition). In fact the three letter code system for all languages, now adopted as ISO codes, was originated by SIL for the Ethnologue. Also as it turns out, project director Jim Mason had once been nursed back to health by an SIL linguistic camp in the middle of the Papua New Guinea jungle while doing his graduate work there. Jim Mason's first step was to make arrangements with SIL to use their materials. We later became one of the very few organizations to ever license this content.

In addition, we had a to choose which portion of the bible to collect. In our research it became clear that portions of the New Testament were likely the most translated, but exactly which sections was sporadic. In addition if we stayed with the Old Testament writings we would cross at least three major world religions (Christianity, Islam, and Judaism), making the choice slightly more even handed. Which story from the Old Testament to collect became the next question. I made a push for the story of Babel (2), as I thought the irony of that story was too good to pass up for this effort. One of the advisers of the project, artist and musician Brian Eno, felt however that this was a story of a vengeful god which he did not like. The other candidate was Genesis (3), the creation story and the opening 3 pages of the Bible. The winning argument for Genesis 1-3 was actually a completely technical one; if we were sending people to archives around the world to scan these documents in languages and scripts that they did not understand, the only reliable text we could grab would be the first 3 pages. Reliably excerpting later sections in the Bible would be nearly impossible without knowledge of each language. The other benefit was that often local creation myths are translated along with the Bible as a way to get people interested in literacy, and we could collect those stories as well.

So we set about collecting Genesis 1-3. Forging relationships with archives that would allow Jim Mason (or a locally hired librarian) to come scan, consumed the beginning of the project. We discovered the fastest and cheapest black and white scanners by trying many of them, and even discovered a funny bit of technology from Hewlett Packard called the CapShare (now discontinued). The CapShare is a hand held scanner that allowed a few pages to be scanned very easily out of a book, without taking it out of the library - or even breaking its spine. Over about 4 months we collected nearly 900 Genesis translations in time for a conference at Stanford that Long Now was putting on about how to create a 10,000 year library in June of 02000. In addition we had a design for the disk that incorporated both the micro-etched pages and a human eye readable design that would hopefully cue the reader to understand what was on the disk, and that it had to magnified. After collecting all

this content, and doing much design work, we found out that the processes for micro-etching and even normal etching were not as far along as we had hoped. It turned out that while etching tens of thousands of pages could easily be done with the FIB technology, they actually had not worked out good ways of etching tens of thousands of *different* pages. The buffer systems created originally for writing relatively simple circuitry would overflow when filled with the massive amounts of text. In addition, to etch the human eye readable content along with the micro-etching presented difficult alignment and fixturing issues. (For instance if you have a graphic etched by one machine next to micro etched pages from another machine, you would need a way to align the disk within a few nanometers in each machine.)

So we did not complete the disk we wanted to for the conference. But we did learn a lot along the way. One of the most important things we learned was that while having the parallel translations was a good start, there are a few other linguistic elements that are almost more helpful if you are to understand a lost language. For instance it took 50 years of concerted effort to translate hieroglyphics after the Rosetta Stone was found. In speaking with linguists we encountered while collecting data we found that elements such as orthographies, phonologies, grammars and word lists gave a lot of information in relatively few pages. In all we ended up with a list of about ten elements that we should be collecting. We also learned that even cataloging and organizing the information once we had it was a huge task, as very few people can even recognize text in more than a few languages.

With further funding from Charles Butcher we set about the larger collection process and built a web site at www.rosettaproject.org where we posted all the data on-line for everyone to comment on and add to. This got the attention of the larger linguistic community, and we eventually received a nearly $1 million grant from the NSF to extend the collection effort to 2,500 languages. With a broader effort and hiring of in house linguists including Dr. Laura Welcher (who later became the project director) we achieved that goal, and found ourselves housing one of the broadest linguistic collections on the net.

During these years of collection several developments occurred in micro-etching. The CTO of Norsam had left to create Serenity Technology using another micro-etching technique. And micro-etching had found a primary market, not in archiving, but in marking gems and diamonds for authenticity. This gave micro-etching a boost to stay alive, but did not help the companies focus on our needs. But in 02008 a new sponsor came to Long Now with a desire to have a disk, and was able to donate enough to project for us to try and finish it. We asked both Norsam and Serenity, and only Norsam came back to us with a quote. We had also redesigned the disk to better fit the technology. The new design would have the human eye readable elements on one side in a high contrast silver etched through a black coating, and all the pages of linguistic data on the other. Each side would be done with a different technology and material, but now they didn't have to interact as in the earlier design.

This design stayed to the 2.8" diameter since we already had so much work done in that size, and had already made stainless steel and glass cases to handle the 3" final disk made from the 2.8" etched pieces, and held in a surrounding bezel.

The human eye readable side would have an outer ring of text, in eight major world languages, spiraling down to about 0.2mm high. Inside of that was a listing of all the languages represented on the disk in order, and grouped by major geographic region. That text was all at 0.18mm high, and was written in a special "engravers font" created for us by a German company URW. This font used as few curve points as possible, and approximated a single stroke instead of the usual outline information. This was important because if the etching machine outlined each character, which is the standard True Type font method, the characters would get muddled by the necessary thickness of the outline. A single stroke font is more like how a you write a character with a pen, utilizing the thickness of the line itself, instead of outlining each character. In the center of all of that is an image of the world. This disk is made out of commercially pure (CP) .035" thick titanium (acquired at McMaster.com) cut into a 2.8" circle by waterjet, and then coated in a matte black oxide finish by Russamer Lab. The material and cutting cost around $100, and having 15 of them coated cost $600 (we made plenty of extras for testing with). The etching was then done by Norsam as part of their quote which is detailed below.

The side of the disk with all the data pages on it was also etched by Norsam. We wanted our pages written so that they ended up in a circle to maximize the page size on the disk. We had also learned along the way that we should not expect to be able to etch more than 15,000 pages, and still have them be easily readable by optics. This meant that we had to edit the archive. We ended up taking the first page of each element of the archive, adding all our phonetic word lists, as well as a few pages at the beginning and end about the project. This left us with about 13,500 pages to be etched. Kurt Bollacker wrote a script that pulled these pages from the archive and organized them into folders, one folder to a row to be etched on the disk. These were also grouped by geographical region and named with a special file name that had meta data about each file (its name, region, three digit code, page number etc). I then fed these through a program, ironically called Debabelizer, to automatically "stamp" each page at the top with the meta data in the file name. The digital stamping was done in another special font called OCR-b which allows for Optical Character Recognition. This would ostensibly allow for a computer to catalog all the data on the disk if it were scanned back in with a microscope. Kurt then took these files and "padded" them with blank files so that if they were organized in a grid, we would end up with all our files in a circle in the center of that grid. These folders were then sent by hard drive to Norsam who etched them into silicon in their FIB machine. Unfortunately they could not maximize it to the disk size as they had said, and we ended up with the pages written smaller than hoped in a 1.9" circle. This silicon master was then transferred to another substrate, and then used as a mold for successive layers of nickel plating to create a pure nickel disk. When they sent us this

disk, it was over-sized, and not centered. We had Jascha at Applied Minds cut it down to a 2.8" circle using a wire EDM machine. For this process Jascha protected the surface of the disk from corrosion during the cutting process by covering the disk with acrylic that was sealed in place with silicone.

The first of these disks was presented to the Wilke Foundation in August of 02008, almost a decade after the project began. You can see a write up of that event and the project by Kevin Kelly. For the etching of both sides of each of the 6 disks Norsam quoted the project at $25,000. The project is not totally complete and it remains to be seen how well they will do on all six disks, and the human eye readable side. I will update this file once it is complete.

1 - Other broadly translated documents: In our research we found that a text by VI Lenin was listed as translated into nearly 1000 languages (but we could never find an archive that had it), and that likely the UN Declaration of Human rights was the second most translated document at around 500 (now freely available on the net). After that Agatha Christie, Jules Verne, Walt Disney and William Shakespeare come in as the next most translated authors.

2 - The Story of Babel: (from Genesis 11)
Now the whole earth had one language and few words. And as men migrated from the east, they found a plain in the land of Shinar and settled there. And they said to one another, "Come, let us make bricks, and burn them thoroughly." And they had brick for stone, and bitumen for mortar. Then they said, "Come, let us build ourselves a city, and a tower with its top in the heavens, and let us make a name for ourselves, lest we be scattered abroad upon the face of the whole earth." And the LORD came down to see the city and the tower, which the sons of men had built. And the LORD said, "Behold, they are one people, and they have all one language; and this is only the beginning of what they will do; and nothing that they propose to do will now be impossible for them. Come, let us go down, and there confuse their language, that they may not understand one another's speech." So the LORD scattered them abroad from there over the face of all the earth, and they left off building the city. more at bible.com

3 - Creation Story in Genesis: (opening lines from Genesis 1 below)
First God made heaven & earth The earth was without form and void, and darkness was upon the face of the deep; and the Spirit of God was moving over the face of the waters. And God said, "Let there be light"; and there was light. And God saw that the light was good; and God separated the light from the darkness. God called the light Day, and the darkness he called Night. And there was evening and there was morning, one day. And God said, "Let there be a firmament in the midst of the waters, and let it separate the waters from the waters." more at bible.com