编码文明

地外文明探索计划(SETI)终于取得了成功!

2022 年，在对宇宙微波频道连续侦听了 38 年后，科学家终于从天鹅座一长串 0，1 信号中成功地解码出地外文明发给全银河系的广播。这一年《三体》电影正在全球热播，造成非常多的科学家和政治家由于黑暗森林理论，反对和外星文明接触。但经过长达一年的辩论，大多数人逐渐认识到这是一个远远超过地球的友好文明，特别在收听到专门对地球坐标发送的问候之后，联系派占据了上风。集中了全球电网 30%的能量并把中国的天眼 500 米射电望远镜改造成发射器后，一束强大的能量从贵州的天坑中直刺苍穹，发送去了人类的应答。

这个应答相当于宇宙 C 级文明的入场券。从此之后，外星文明和地球的通讯繁忙起来。预想中的 265 光年造成的延时神秘地并不存在，地球发出的电文往往在瞬间就能收到回复。终于地球在 2024 年迎来了迎接宇宙使者的那一天。

由于宇宙使者要求造访中国国家图书馆，联合国外星接待委员会又希望把这一事件对地球居民日常生活的干扰减小到最低，于是我这个刚毕业的图书馆管理员就意外地肩负起了接待重任，不过好像中国历史上图书馆管理员一向是个很重要的角色。

晚上十点闭馆后，我在忐忑不安的心情中等待了 30 分钟。随着一阵蓝光闪过，宇宙使者就突然降临在我的面前。尽管设想了无数次外星人的形象，眼前这个美丽的女子还是让我大吃一惊。以前所有见过的美女在她面前都要黯然失色，一层淡淡的光晕完美地勾勒出她瀑布般的黑发，星眸盖过了窗外北京霓虹的夜色。她手上拿着一根银色的仙女杖似的东西，更显得优雅。

“没有吓着你吧？”嫣然一笑让人如沐春风，“我用了你们星球青春期女性相貌的平均值，据说这样对你们星球上的另一个性别最有亲和力。”

“没有”，我仿佛又回到了在大学面对我那个校花学姐的时刻。但几年来在图书馆书香的熏陶下，我多少也变得有些老成。肩负着人类的使命，我努力平复着心跳，用尽可能平静的声音说：“欢迎来到地球，欢迎来到北京，怎么称呼您？有什么可以帮助到您的吗？”

“你叫我琳吧，其实你什么也不用做，我来地球只是收集一下你们文明发展的信息，大约地球时间 30 分钟就好了。这些植物纤维上印着碳元素标志的东西都是你们知识的载体吧？”琳说完后一挥仙女杖，一团不易察觉的蓝色雾气向书架水银泻地般涌去

“中国国家图书馆馆藏宏富，品类齐全，古今中外，集精撷萃。馆藏文献超过 3500 万册件并以每年百万册件的速度增长。馆藏总量位居世界国家图书馆第七位，其中中文文献收藏世界第一，外文文献收藏国内首位。国家图书馆馆藏继承了南宋以来历代皇家藏书以及明清以来众多名家私藏，最早的馆藏可远溯到 3000 多年前的殷墟甲骨。珍品特藏包含敦煌遗书、西域文献、善本古籍、金石拓片、古代舆图、少数民族文字古籍、名家手稿等 280 余万册件。“敦煌遗书”、“赵城金藏”、《永乐大典》、文津阁《四库全书》被誉为国家图书馆“四大专藏”。”对于国家图书馆我如数家珍。“我们还收藏了丰富的缩微制品、音像制品，还建成了中国最大的数字文献资源库和服务基地，数字资源总量超过 1000TB，并以每年 100TB 速度增长。实施“中国记忆”项目，围绕中国现当代重大事件、重要人物等专题采集口述、影像、音频等文献史料。”

墙上硕大的标语写着：“书籍是人类进步的阶梯”，人类的知识在握，我每天看到这个标语都感到由衷的自豪。

“1000TB，也就是 8x10^16 比特”，琳回答道：“这是你们度量信息的单位吧？”

“是的，我们现在已经完全数字化,所有文字资料都按 Unicode 编码，可以进行检索。”这些年图书馆的数字化建设进展很快，我作为刚毕业的图书专业毕业生，领导给我压了不少担子，也学到了不少新知识。

“听起来不少。”琳眨了眨她美丽的大眼睛，“我来看看究竟有多少信息熵。”

“信息熵？”我还是头一回听说这个名词。

“是啊，这和你们发明的长度单位米，质量单位公斤一样，是衡量信息量的一个单位”。琳指着一排电子类图书的书架说：“听说过信息论的创始人香农吧？”

这个名字以前倒是听过，主要是好记，香喷喷的农民，和北京街头看到的农民工正好相反。我回答到：“听过，但我不是电子专业的。“

“其实很好理解，话语的信息量取决于这个话语描述的事情的发生概率，概率越小，信息量越大，概率越大，信息量就越小。“

蓝色雾气笼罩了这一层的各个书架，琳开始读取书架上的信息了。速度很快，这团雾气弥漫开来，很快充满了房间，又向楼层外涌去。

“比如说，我告诉你你们星系的那颗恒星，你们叫太阳的那个，明天还会在东方升起，这几乎是必然的事件，这句话的信息熵就很小。如果我说明天太阳会从西方升起，这个概率就非常小，而这句话的信息熵就很大。“

“懂了“，有个大美女和你讲科学知识就是容易学。

收集信息的过程好像并不需要琳的干预，她接着说：“你们的文字每个字的信息熵是不同的，用英语来说，26 个字母中 e,t 出现的概率最高，它们的信息熵就最小。较少出现的字母比如 z 信息熵就大。单个英文字母的信息熵平均为 3.9 二进制位，也就是 3.9 比特。单个汉字的信息熵是 9.56 比特。 “

“对于信息熵小的字母，我们可以用很少的位来编码。比如你们海上通讯用的莫尔斯电码 e 就是一短声：滴，t 就是一长声：答，这样发送起来就快很多。少用的 z 就是“答答滴滴”，发送起来就比较费时。但由于电文中通常 e,t 多，z 少，整个电文发送起来效率还是比平均分配码长到每个字母高。Unicode 每个字符都用 16 比特编码，是一种非常低效的编码方式。“

“明白了信息的熵，接下来就要用到熵编码了。”琳继续向我科普。“刚才我讲的信息熵是无形信息的单位，无论用哪种编码办法，比如用中文，英文，文言文，编码出来的位数一定会比信息熵多，不可能比它少。”。

“那当然，物品加上包装一定会比原来重，我上周在淘宝上卖了 5 斤核桃，回来连包装才 4 斤半。。。”我心情放松了些，有些啰嗦了。

“你们地球上一些天才早就发明了比较好的编码方法，比如霍夫曼编码，他二十五岁为了免于考试，就写了课堂作业代替，结果就发明了这种方法。原理上就是用比较短的码来表示出现概率最高的符号，长一些的来表示概率低的符号。这种编码你们已经用在 JPEG,MPEG 上用来编码图像和视频了。”

“我们还有压缩软件，你收集的信息要不要压缩一下再带回去？”我贴心地问。

“你们的压缩软件用到了更好的算术编码，这是目前你们最接近于信息熵的编码方式。

比如我们要把“ARBER”这个英文单词编码，注意到 R 在这当中出现了两次，这样每个字母出现的概率就是：

Symbol	Times	*P*
A	1	0.2
B	1	0.2
E	1	0.2
R	2	0.4

我们在 0-1 的区间上给每个字母分配一个区间，**A 占用 0-0.2，B 占用 0.2-0.4，E 占用 0.4-0.6，R 占用 0.6-1.** 注意由于 R 出现的概率大，所以给它安排的区间也宽。

当我们看到第一个 A,我们选择 0-0.2 区间，然后我们再把 0-0.2 区间按照上面的概率重新划分，新的区间就是：

A:0-0.04

B:0.04-0.08

E:0.08-0.12

R:0.12-0.2

第二个要编码的字母是 R,所以我们取 0.12-0.2 区间，然后再把 0.12-0.2 区间按字母概率划分。

由此类推，下一个字母是 B，我们取 0.136-0.152 区间。

再下个字母是 E，我们新的区间是：0.1424-0.1456。

最后一个字母是 R，我们最终的区间是 0.14432-0.1456。我们在这个区间随便取一个数比如 0.145， 这个数就代表了”ARBER”的全部信息。

解码首先看到 0.145 这个数是在 A 的 0-0.2 区间，所以第一个字母就是 A.

我们从 0.145 中减掉 A 的起始位置 0，再除以 A 的概率 0.2，得到(0.145-0)/0.2=0.725, 落在 0.6-1 的 R 区间，所以第二个字母是 R.

知道第二个字母是 R，我们把 0.725 减掉 R 的起始点 0.6 再除以 R 的概率 0.4，我们得到：（0.725-0.6）/0.4=0.3125， 这个数落在 B 的区间，所以我们知道第三个字母是 B.

用同样办法，我们得到第四个字母数字是：（0.3125-0.2）/0.2=0.5625,落在 E 区间，第五个字母数字是：(0.5625-0.4)/0.2=0.8125，落在 R 区间。这样我们就把“ARBER”还原了。“

“真神奇！“

“霍夫曼编码固然不错，但最少也要用 1 个比特。如果你要编码 800 个 A，再加上 200 个 B,用霍夫曼编码的话，用 0 代表 A，1 代表 B，编码结果是：0.000……000111…111 (800 个 0，200

个 1)，你还是要用 1000 个比特。“

“但是如果是算术编码呢？因为 A 的概率是 0.8，所以算数编码会使用区间 [0, 0.8) 来编码 A，800 个 A 则会形成一个区间 [0, 0.8^800=2.964760347899781341208136941058e-78)，显然这个区间比 [0, 0.5^800=1.499696813895630954817644437628 1e-241) 位数少得多，也就是说 800 个 A，哈夫曼编码用了整整 800 个 0，而算术编码只需要不到 800 个 0，更少的比特数就能表示。“

“所谓编码就是用尽可能少的比特数来表示信息，最后不管是一个十进制数还是二进制数都一样，对吧？”我听得津津有味。

“是的，你很聪明！”琳又给了我一个微笑，我脸上不由得有些发烧。

“您刚才说中文字符每个字的信息熵比英语高得多，这是您选择来中国国图收集信息的原因吗？“我插话道。”中文是最精练的语言，我们联合国有五种官方语言，通常中文版的那份文件是最薄的。。。“我不禁有些自豪地说。

“如果按照每个字去编码，那是个很笨的办法。“琳狡黠地眨了下大眼睛。”不管是中文，英文，你们都有词的概念。比如说，如果前面六个字是“中国国家图书“，那后一个字应该是啥？”

“馆吧？，当然也有可能是“出版社”等等。“我接着说。

“是的，你可以看到字之间是有非常强的关联的，用词来编码，效率会高很多。“琳走到一边的工具书架上，抽出一本汉语大词典，”你看，你们也早就用上这种办法来表达你们的意思了。“

“我们祖先还创造了一种文言文，编码效率应该更高！“，我不禁想在琳面前多表现一下。

“是的，从编码的角度，文言文确实是一种神奇的高效系统，而且你们连标点符号都省略了。“琳掩嘴一笑：“但你们的祖先通过喉部机械震动来交流，也就是讲话，可不同于书面语言文言文。地球上的所有生物由于受声波交流方式的限制，人类最快每秒钟也就是几十个比特吧，地球上各种语言在这一点上都差不多。比如你看不起的英语，单个字或词的信息熵低于汉语，但语速就比较快，每秒钟可以说更多的字和词，这样交流的效率也是差不多的。”

确实，如果我们今天用文言文讲话的话，肯定讲得累死，我不禁恍然大悟。

“我看到你们有人在做脑机接口了，“琳看了一下一旁杂志架上特斯拉汽车的标记，”等你们交流的速度达到千比特，兆比特，你们会打开一个全新的世界的。“

汉语书面和口头表达是两套不同的编码系统，学习起来比较困难，文言文也就慢慢淘汰了，汉语的书面表达也日益口语化。拼音文字用同一套编码系统涵盖书面和口头交流，是很大的优势。

蓝色的雾气开始从图书馆各个角落流淌回来，向仙女棒上凝结。

“你们地球人还是很聪明的，比如我刚才听你说“国图”来代替“中国国家图书馆”，这就是提高交流效率的一种方法，你们学术上叫字典编码。“琳看我有些失落，安慰道。

“是啊，我们每年创造很多新词呢，比如“内卷”，“996”。。。“这些概念有些负面，违背了我想努力在琳跟前留下好印象的初衷。

“奇怪的是你们也浪费了很多字去编一些无意义的东西，比如“会议没有不隆重的,闭幕没有不胜利的,讲话没有不重要的,鼓掌没有不热烈的”。。。“琳看着另一排书架，上面都是些时政文献

之类的东西。

我不禁有些汗颜，这些都是些空话套话，但每天好像也不得不说，本来就每秒就只能交流几十比特，还要浪费很多在这些空话套话上，唉。我想岔开话题，就问道：“您读取了这些信息怎样带回去呢？”

“你看，“这时蓝色的雾气在仙女棒上越聚越浓，最后在中间某个位置形成了一个漂亮的圆环标志。”如果我这根棒子的长度为 1 的话，我在中间 0.51389077845…的地方做了个标志“。

“这串数字有啥意义吗？“

“这串数字就是你们地球文明所有知识的编码。“我选了个和你们星球颜色相同的蓝色，你喜欢吗?” 琳在挥舞着那条仙女棒，我看到棒子上还有着数不清的各色的圆环标志。

“你们老祖宗不是说过：“一尺之棰，日取其半，万世不竭”吗？不要被组成这根棒子的原子尺寸限制，也不要被普朗克长度限制，你们猜测到宇宙是美妙的琴弦，但离真相还早呢。”在采集完地球信息后，琳仿佛立即对我们的文明了如指掌了。

“一切语言都是重复，一切交往都是初逢，这两句诗挺有意思的，再见！”。琳挥舞着仙女棒慢慢地消失在空气之中，我脑海里还在萦绕着那串长长的数字，所有的人类活动，知识，文明的印记只是那一圈淡淡的蓝色的光环。