

老子道經
河上公章句第一
體道第一

道可道 謂經術政教之道也 非常道 謂自然長
名可名 謂富貴尊榮 非
有欲以觀其妙 妙要也人常能
有欲以觀其微 微歸

騷經第一 離騷一
離騷經者屈原之所作也屈原
同姓仕於懷王為三閭大夫三閭
王族三姓曰昭屈景 戰國策楚有
子瑕食采於屈因氏焉屈重屈道
平並其後又云景氏有景差至
中 屈原序其譜屬率其賢良以
則與王圖議政事決定嫌疑以
對諸侯謀行職脩王

初五
天地然後有萬物
然後有夫婦有夫婦
有君臣有君臣然後
有所錯
曰言咸卦之義也咸
與夫婦之象莫美
故夫子殷勤

48072
蓮華經者統諸佛降靈之本致也蘊結大
出彼千齡東傳震旦三百餘載西晉惠帝永
康年中長安青門燉煌菩薩三法護者初翻此
經名正法華東晉安帝隆安年中後秦弘始丘
慈沙門鳩摩羅什次翻此經名妙法蓮華隋氏
仁壽大興善寺北天竺沙門闍那笈多後所翻
者同名妙法三經重沓文旨互陳時所宗尚皆
弘秦本自餘支品別偈不無其殊具如序曆故

傳杜氏序
魯史記之名也記事者以事繫日以日繫
繫時以時繫年所以紀遠近別同異也故
記必表年以首事年有四時故錯舉以為
也周禮有史官掌邦國四方之事達四
春官小史職曰掌邦國之志內史職曰凡四方之事書內
外史職曰掌四方之志掌達書名於四方今杜氏序云
者掌邦國取小史職文四方之事取內史職文周禮言
侯之國也國有四表故言四方云凡四方之事書內
來告內史讀以白王也告王之後則小史主掌之
云讀四方之事書其實內史掌實內史所掌
然則內史小史既主國內及主四方來告
告而書是也杜此序又云達四方之

今秘閣古文通止

——AI时代古文整理新尝试

宋杰

2023年10月5日

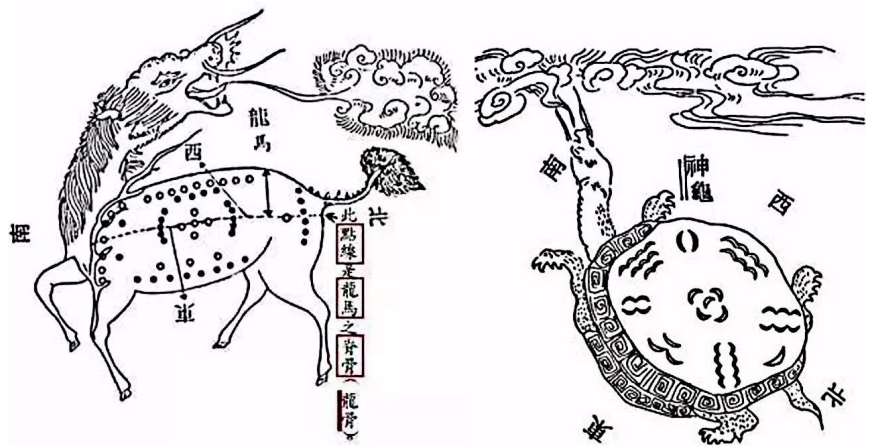
終南山釋 道宣述

古者伏羲氏之王天下也，始画八卦₂，造书契₃，
以代结绳₁之政，由是文籍₄生焉。

——古文《尚书·序》 孔安國

让书写在古籍里的文字活起来。

—— 习近平



AI语言大模型时代古文的尴尬处境

- 英文主导地位难以撼动
- 古文字Unicode编码不完备
- 字形信息完全被抛弃
- 繁简混用、古今混用

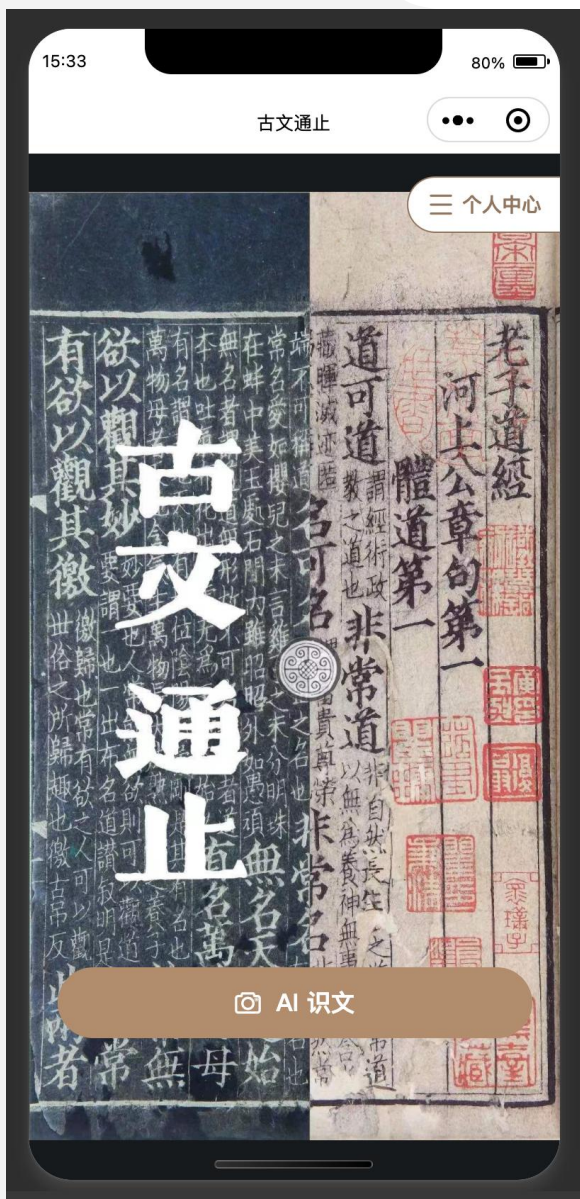
根本原因在于**编码本位**的汉字数字化方法天生不足以表达**高维度信息**

古文活化的主要障碍

- 字词生僻（异体字、通假字、避讳字）
- 语义晦涩（分词断句、古今异义）
- 思维模式差异（编码思维与图像思维）
- 古文字基础数据集共享模式缺失

AI时代呼唤古文整理新范式的出现





古文通止的设计理念

- 基于IIIF（国际图像互操作）的**图像本位**而不是**编码本位**的古文数字化基座
- 整合计算机视觉与自然语言处理的端到端的古文字手写体OCR算法模型
- 以字词释义为核心、一体化运用知识图谱与大语言模型LLM（ChatGPT与文心一言）的古文翻译
- 解释专有名词时，灵活协调与AI之间人机交互模式的古文Copilot
- 基于Git与TEI的创新型古文版本管理，并以AI机器人实现自动审核
- Web3与DAO加持下的古文价值注入与参与者确权



古文通止的原则与步骤

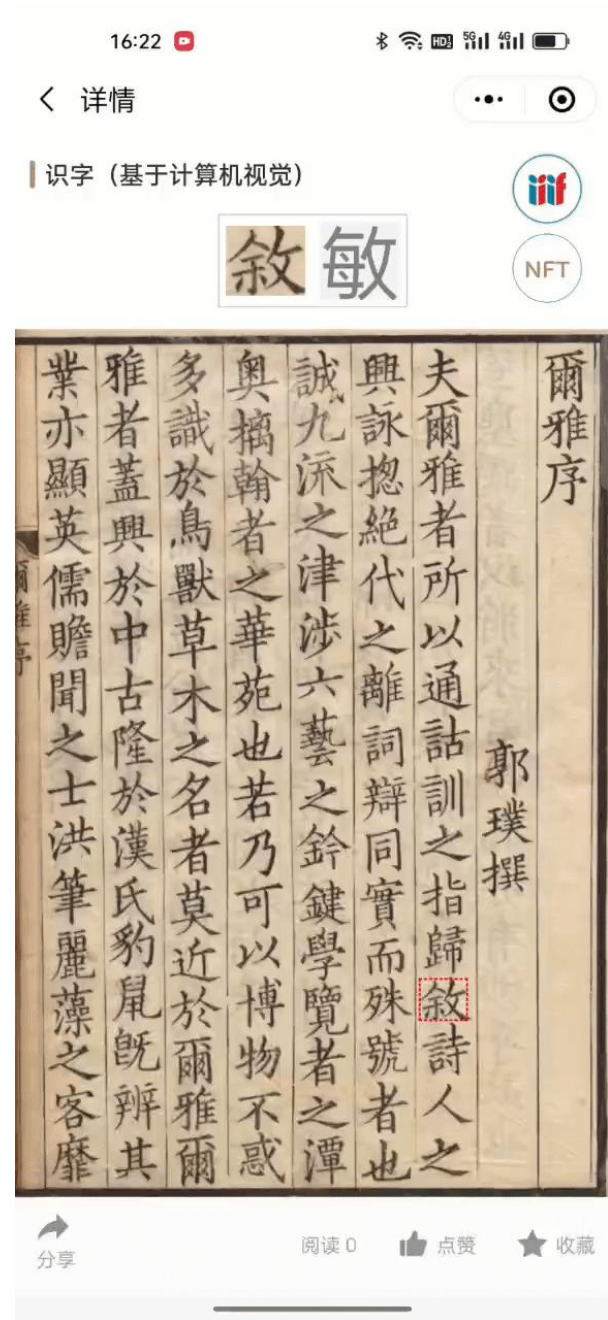
信 —— 字形辨识

达 —— 字词释义

雅 —— 全文翻译

信 —— 字形辨识（一）

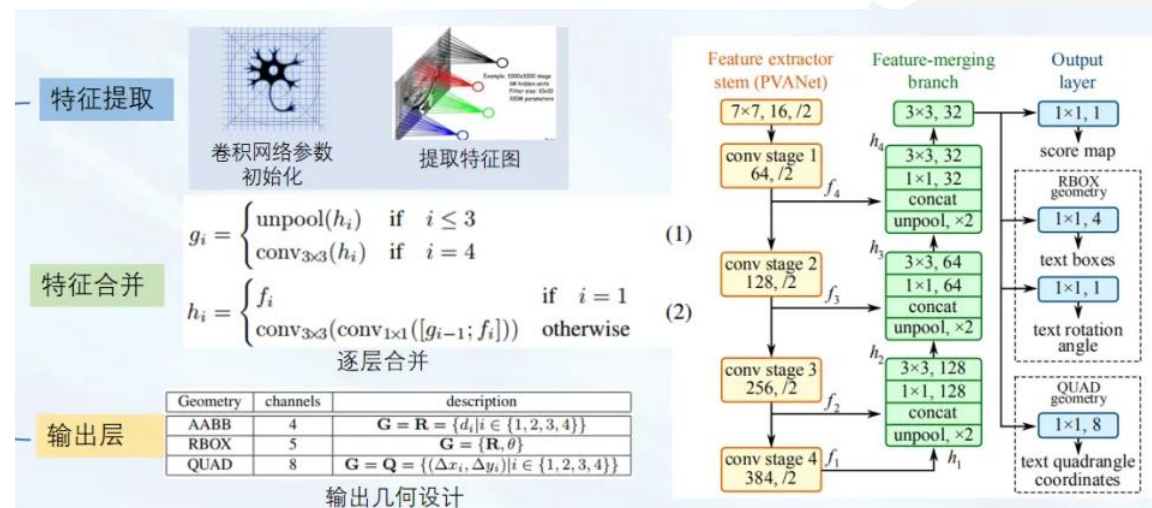
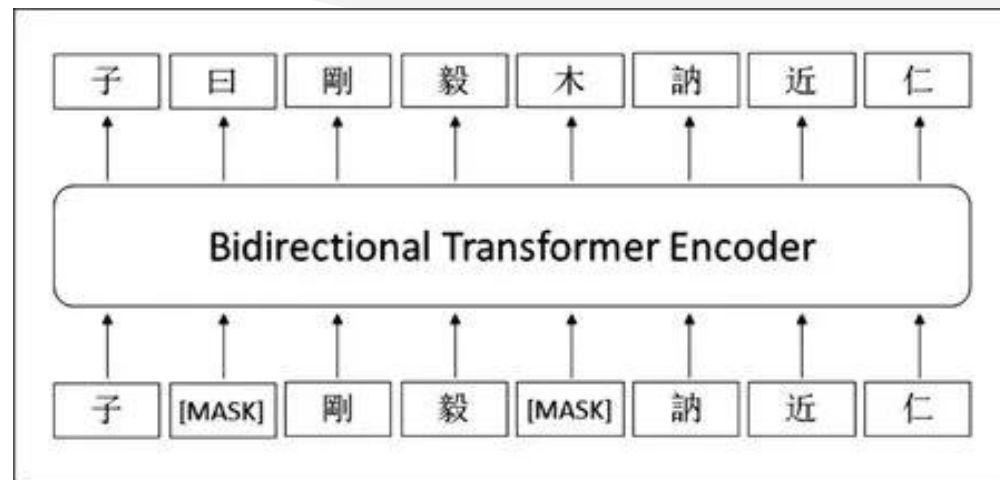
- 面向印刷体OCR软件无法有效识别古代文字
- 中文汉字OCR需要大规模文字图像工程加上识别算法创新
- 基于众包的移动式OCR校正软件有利于汉字图像数据的累积
- 基于IIIF（国际图像互操作框架）的数字底座是实现高质量汉字手写体OCR的工程化基础

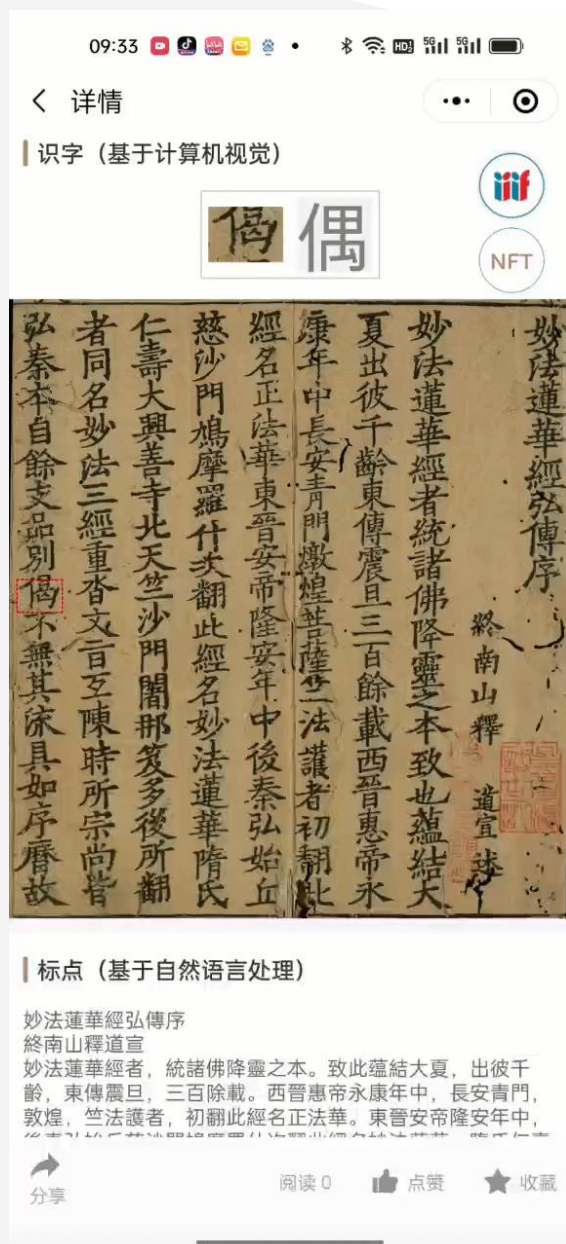


信——字形辨识（二）

SikuBERT是以四库全书为语料以BERT模型进行培训的大语言模型。与生成式大语言模型如GPT相比，BERT模型更擅长做语义理解任务，

比如文本摘要与完型填空。在古文OCR算法中加入基于单字识别置信度的字符校正，能够有效提高算法的整体识别率。





达——字词释义（一）

- 古文中存在大量人名、地名以及历法、官职、典故等现代读者难以理解的专有名词
- 古文Copilot方式以人为主、AI为辅的方式极大提高工作效率
- 权威参考：上海图书馆人名规范库、CBDB中国历代人物传记资料库、CHGIS即中国历史地理信息系统、浙大智慧古籍平台
- 数据补充：ChatGPT与文心一言等大语言模型



达 —— 字词释义（二）

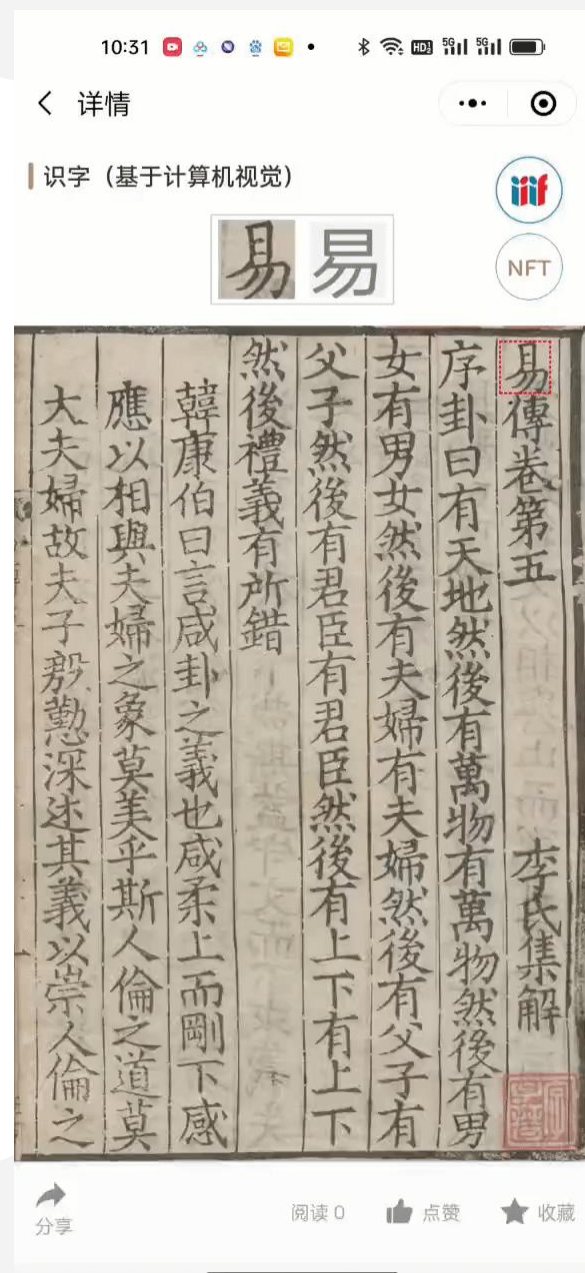
- 释义有分歧的古文可以使用GitHub的版本控制系统进行管理，并由**古文通止机器人**评论——用魔法打败魔法
- AI时代实现**乾嘉学派**的学术理念
- 将字词释义与原始文本以**TEI**格式保存，古文原始图像以**IIIF**形式发布，从而积累高质量数据集
- 古文识读专家的专业释义，并过合理分类整理形成**古文数智人**

雅 —— 全文翻译（一）

语言大模型缺少领域知识，特别是中国古代文化知识是必须克服的一道难题

整合知识图谱与语言大模型需要的是对于知识表示与知识计算的深刻理解

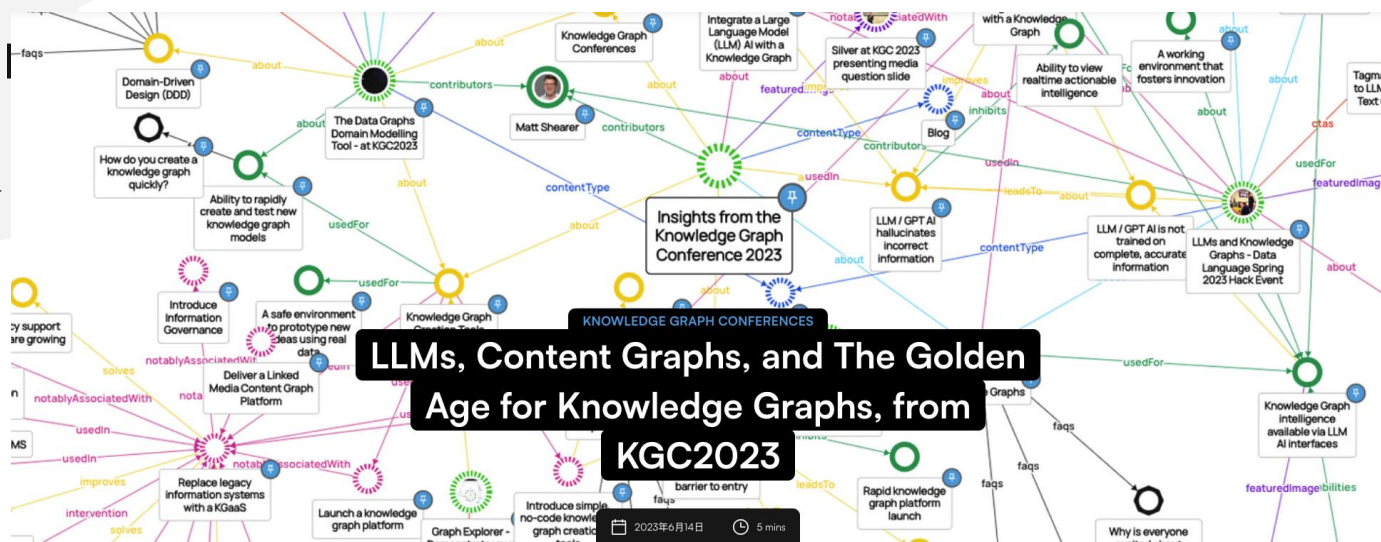
在知识表示和知识计算两个任务使用**显式知识**和“**参数化**”知识的比较中，知识表示更加偏向**表达性**和**判定性**的权衡，而知识计算更加偏向**精确率**和**召回率**之间的权衡

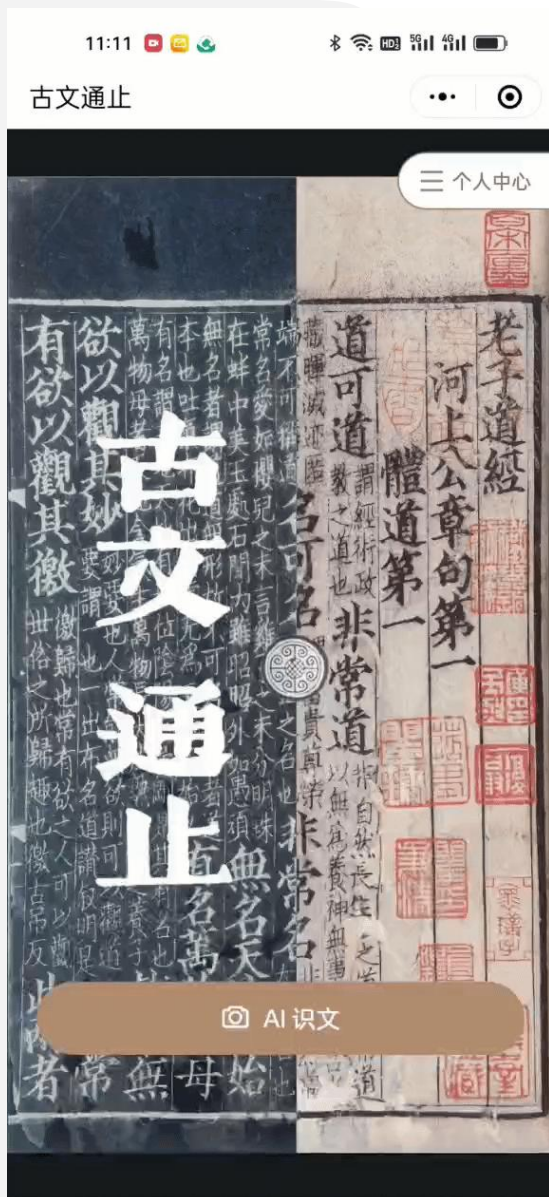


雅——全文翻译（二）

为大语言模型增加领域知识的实践:

- GPT-3.5 Turbo支持自定义微调
- GPT Function calling 实现对于专业词汇解释地无缝嵌入
- Nebula Graph Store LlamaIndex 实现知识图谱的整体嵌入
- Langchain加vector Store 方式为LLM补充领域文本知识
- Semantic Kernel SDK 提示词工程化



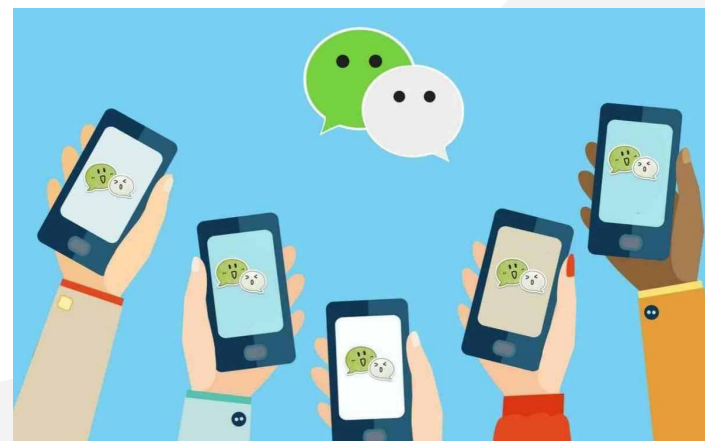


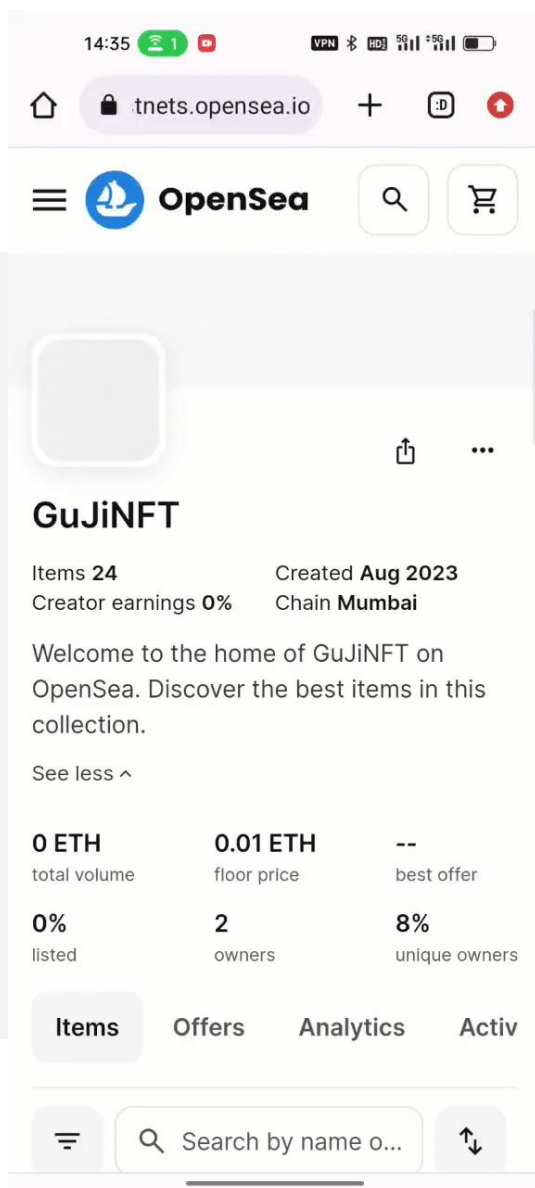
微信小程序传播

社交网络如微信小程序，使用便捷，覆盖面广便于转发点赞等二次传播

有效管理小程序的用户，并对专家用户与普通用户进行区别有利于提高生成内容的质量

对于高质量的古文翻译内容，除了进行流量扶持，更可以使用NFT进行价值重估，产生正向激励





Web3与古文DAO

目标：以平等开放的精神促进古文数字化资源跨机构跨国界流动，激励研究人员以高效、协同地方式共同开发数字化古文宝藏

图档博机构：以智能合约方式面向全球征集原始数据集，提出任务/项目，并设计相应利益分成

机构/独立学者：承接来自主导机构的文化开发任务/项目，在无需担心利益纠纷的前提下，尽全力完成任务/项目

IT供应商：利用以太坊上丰富的dApp以及IPFS/Filecion的分布式存储，为机构及学者提供技术支持并通过智能合约以代币方式结算

前提：符合IIIF标准的全球化分布式数字基座，如易图，是古文DAO的基础保障

古文观止的IIIF标准Web3分布式数字基座——易图

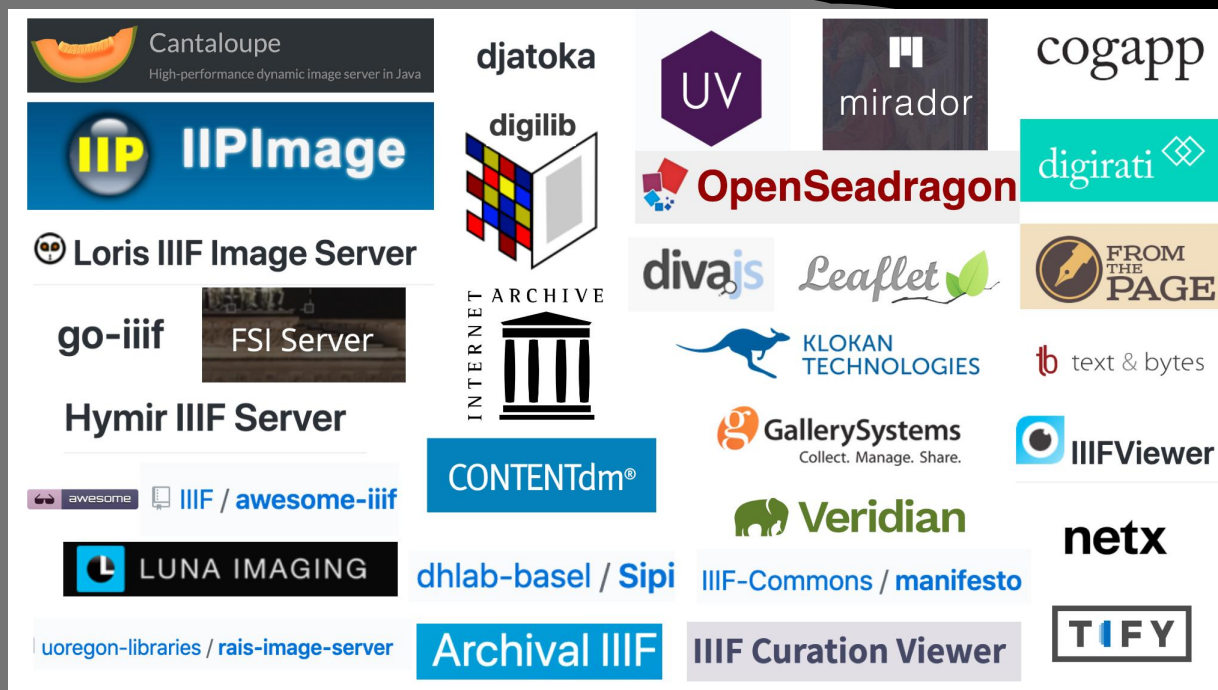


2022年6月在哈佛召开的IIIF年会上发布易图CLI



International Image Interoperability Framework

国际图像互操作框架



A Community

一个社区

that develops shared APIs,

制定接口

implements them in software,

实现软件

and exposes interoperable content

呈现可交互内容

Museums / Galleries

British Museum (大英博物馆)
National Gallery of Art
The J. Paul Getty Trust
The Walters Art Museum
Yale Center for British Art
Et al.
... more

Aggregators

ARTstor
CONTENTdm
DPLA
Europeana (欧洲记忆)
Internet Archive
Wikimedia Foundation
... more

State / National Libraries

Austria
Bavarian State
British Library (大英图书馆)
Denmark
Egypt
France
Israel
Japan
Moravian Library
New Zealand
Norway
Poland
Scotland
Serbia
Wales
Qatar
United States (LoC)
Vatican (梵蒂冈图书馆)
... more

Universities and Research Institutions

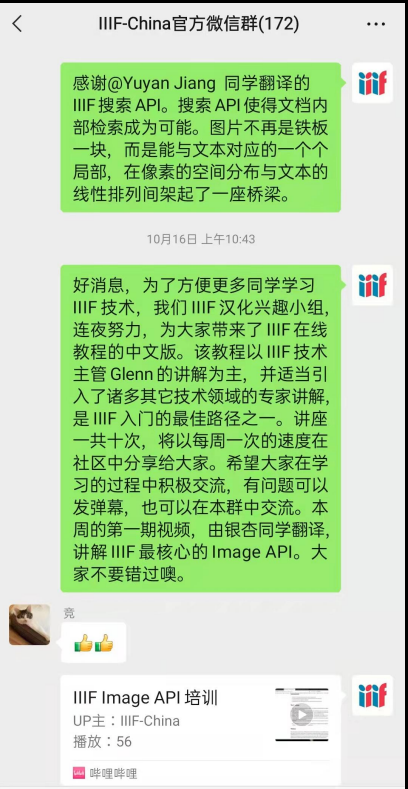
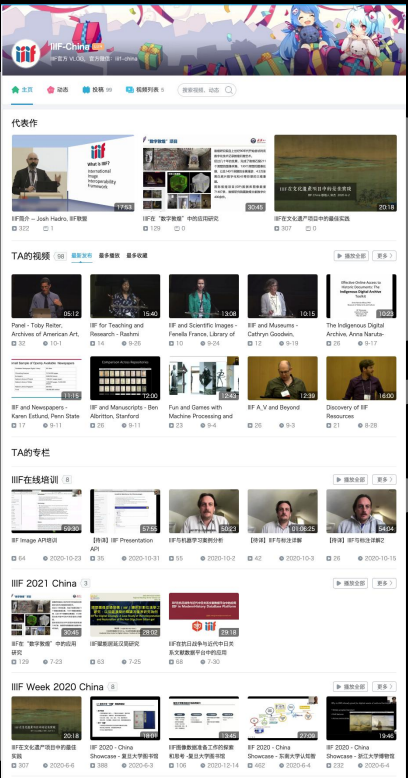
Cambridge (剑桥)
Cornell
Ghent
Gottingen
Harvard (哈佛)
Oxford (牛津)
Princeton
Stanford (斯坦福)
Edinburgh
Toronto
Wellcome Trust
Yale (耶鲁)
MIT (麻省理工)
... more



2017年IIIF梵蒂冈年会

IIIF—China IIIF 中国社区

参与机构：
中国社科院
台北中研院
敦煌研究院
复旦大学图书馆
香港中文大学图书馆
武汉大学数字人文中心
浙江大学艺术博物馆
中国美院
东南大学认知智能研究所
...更多



团队介绍



云技术架构师

复旦MBA

今秘阁创始人

IIIF推广大使

后端开发



前端工程师

界面设计

体验设计

美术设计

小程序开发

编后感想

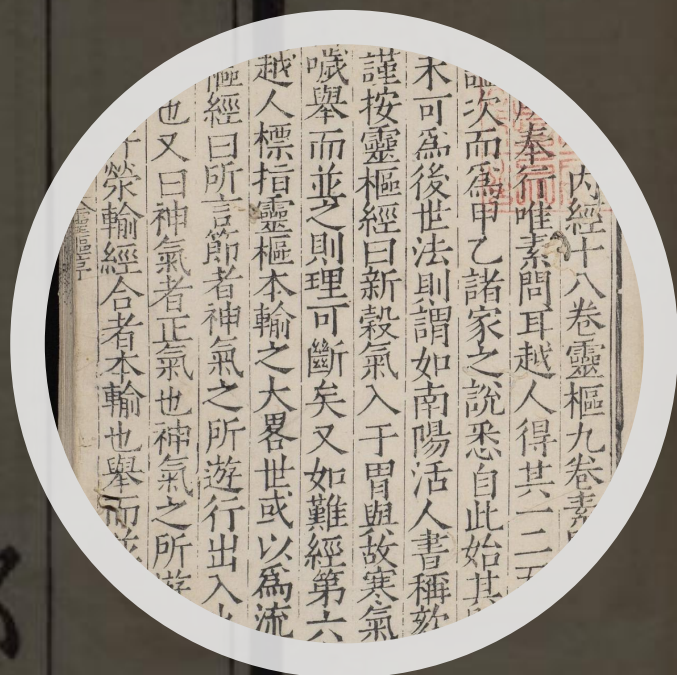
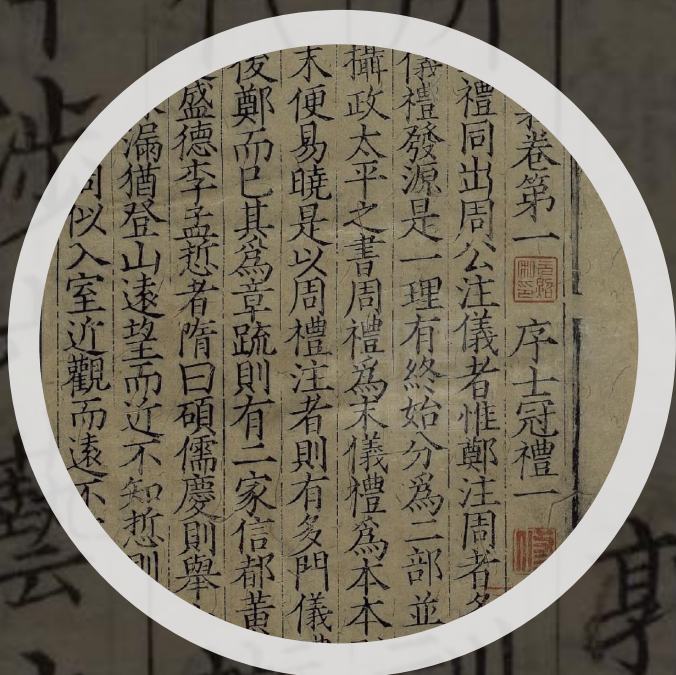
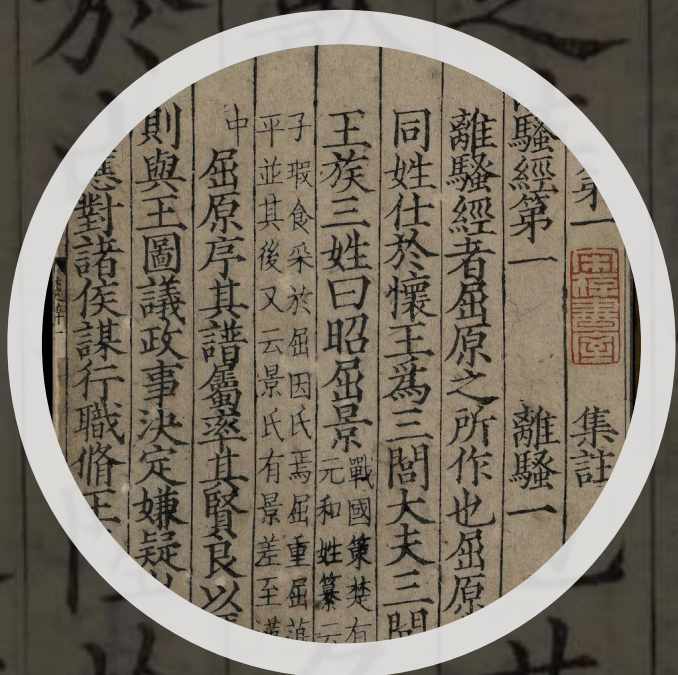
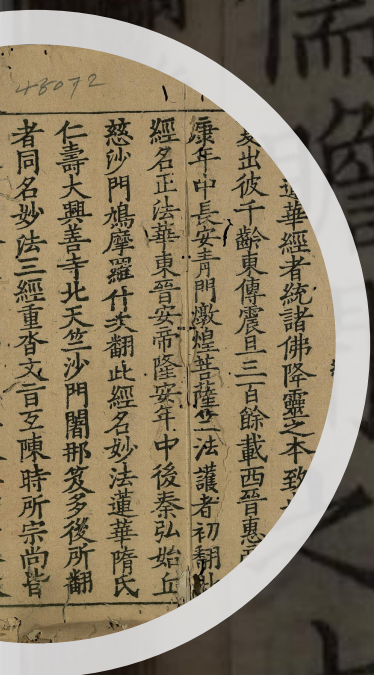
缘起于去年收到的前往**大英图书馆**研究敦煌**妙法莲华经**手写体识别的奖学金。

图像本位的汉字保存方式将是汉字数字化未来的趋势。

古文OCR是图像文字工程，需要**IIIF数字底座**、AI人机接口及图像文字双向检索能力。

之后开放数据能提供知识图谱更多全文访问能力，则与大语言规模整合效果更佳。

Web3 DAO能够协调平台、机构以及专家之前的利益分配，值得深入研究。



IIIF—China
微信号



IIIF 中国推广
大使 宋杰

感谢聆听