

OpenKG进展报告

— cnSchema 2.0 & OpenBase

2018-08-17
王昊奋 丁力
天津



OpenKG技术体系

cnSchema (OpenBase Ontology)



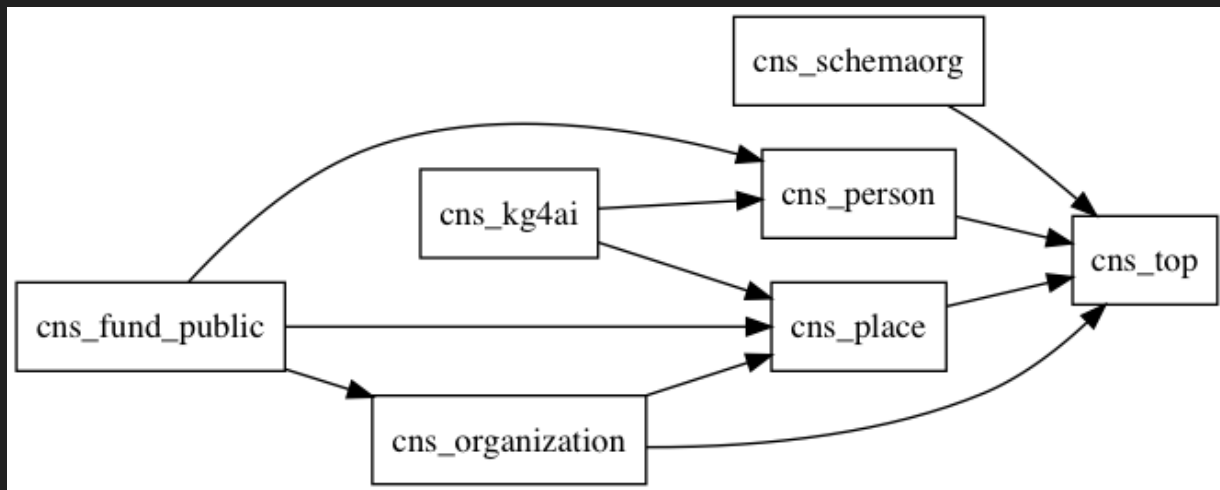
- 社区维护的开放知识图谱标准
- 应用于知识图谱、聊天机器人、搜索引擎优化等
- 提供可供参考的扩展的数据描述和接口定义标准

- 知识抽取、编辑、挖掘、融合
- 知识的更新
- 众包机制审核知识图谱
- 专家委员会的验收
- 自动化构建+人工审核 迭代
- 开放开源cc0协议，开放使用
- 工具平台

- 图谱转化为bot的平台
- 可选择算法，模型，词典和例句来生成bot

cnSchema

cnSchema 面向中文信息处理，为开放中文知识图谱的建模、生产、推理、学习、交互以及应用落地提供可解释、可迭代、可复用的数据接口定义与标准。



cnSchema 1.0: 中文知识图谱的Schema cnSchema.org

[cnSchema](#) [项目](#) [文档](#) [词汇表](#) [关于](#) [OpenKG.CN](#)

CreativeWork

Canonical URL: <http://cnschema.org/CreativeWork>

[Thing](#) > [CreativeWork](#)

【作品】最通用的创作工作, 包括书籍, 电影, 照片, 软件程序等 The most generic kind of creative work, including books, movies, photographs, software programs, etc.

Linked to schema.org: <http://schema.org/CreativeWork> Usage: Between 250,000 and 500,000 domains

Linked to Wikidata: <https://www.wikidata.org/wiki/Q17537576> (creative work)

Linked to Wikipedia: https://en.wikipedia.org/wiki/Creative_work

属性	值域	定义
Properties from CreativeWork		
about	Thing or CreativeWork or Event or CommunicateAction	【关于】事物的关于信息 The subject matter of the content.

概念定义: 中文翻译+WIKI链接。基于[schema.org](#) 3.2 核心词汇, 由清华大学, 浙江大学, 复旦大学, 东南大学, 海知智能, 以及社区志愿者共同完成

[cnSchema](#) [项目](#) [文档](#) [词汇表](#) [关于](#) [OpenKG.CN](#)

[AutoPartsStore\(汽车配件店\)](#)[ClothingStore\(服装店\)](#)[ConvenienceStore\(便利店\)](#)

cnSchema.org 的 Autocomplete 词汇搜索服务

x

[EntityCoreAPI](#)[Get Entity](#)[Lookup Entities](#)[EntityIndexAPI](#)[Cypher Graph Query](#)[ElasticSearch Query](#)[EntitySyncAPI](#)[Batch Update](#)[Delete an entity](#)[List Entities](#)[Update/Create an entity](#)

EntityCoreAPI - Lookup Entities

redis lookup service. 按名字 获取一组实体简略信息 (@id, @type, name, entityScore).

[POST](#)
`entities/`

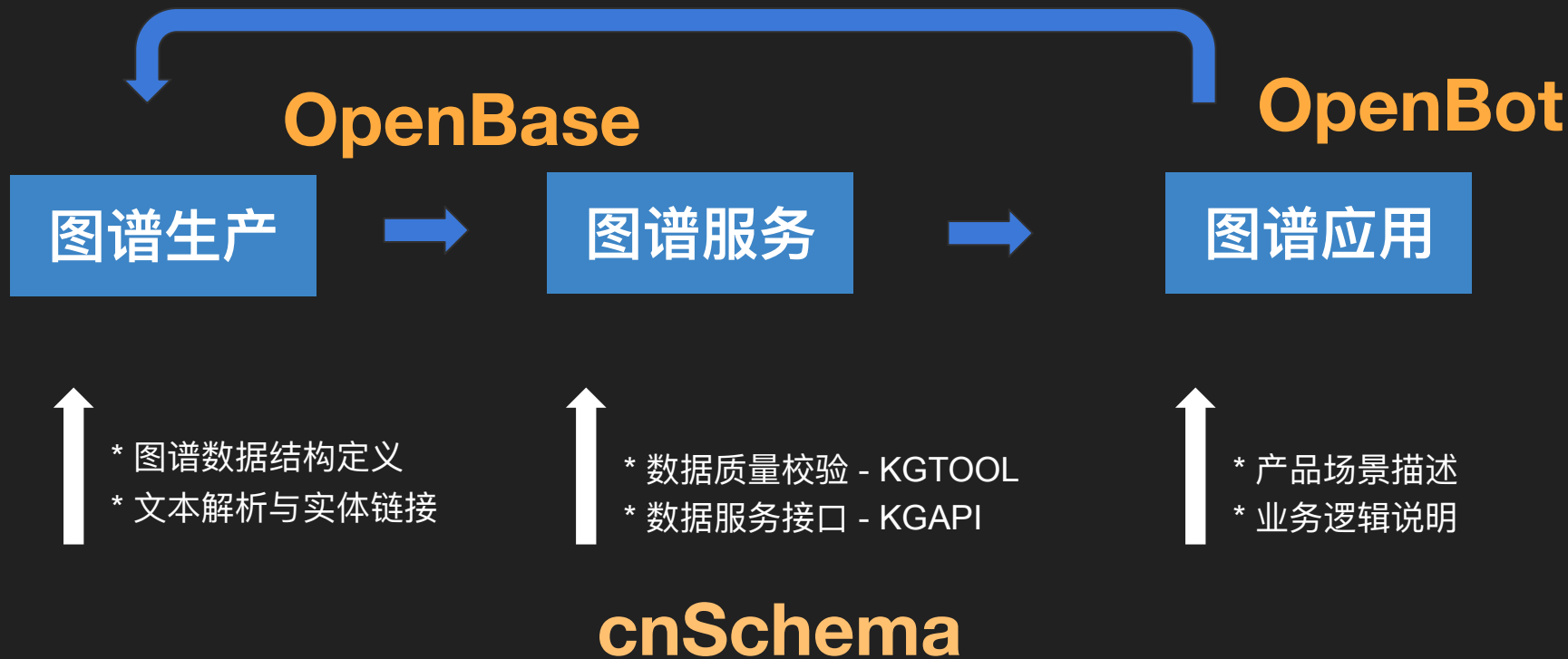
Parameter

Field	Type	Description
names	String[]	必须 list of Entity name, match whole word.

```
{  "names": [    "华子",    "唱歌"  ]}
```

KGAPI接口规范

cnSchema2.0: 开放中文知识图谱的模板化解决方案



百科

北京大学

进入词条

[běi jīng dà xué] 

北京大学

编辑

Peking University

985工程211工程研究生院

北京大学（*Peking University*）简称“北大”，诞生于1898年，初名**京师大学堂**，是中国近代第一所国立大学，也是最早以“大学”之名创办的学校，其成立标志着中国近代高等教育的开端。北大是中国近代以来唯一以国家最高学府身份创立的学校，最初也是国家最高教育行政机关，行使**教育部**职能，统管全国教育。北大催生了中国最早的现代学制，开创了中国最早的文科、理科、社科、农科、医科等大学学科，是近代以来中国高等教育的奠基者 ^[1-3]。

1912年5月3日，**京师大学堂**改称北京大学校，**严复**为首任校长 ^[4]。1917年，**蔡元培**出任校长，“循思想自由原则、取兼容并包之义”，把北大办成全国学术和思想中心，使北大成为**新文化运动**中心、**五四运动**策源地。1937年抗日战争爆发，北大与**清华**大学、**南开**大学南迁长沙，组成**国立长沙临时大学**。不久迁往昆明，改称**国立西南联合大学**。1946年10月在北平复学 ^[5]。

北大由教育部直属，中央直管**副部级**建制，是国家**双一流** ^[6]、**211工程**、**985工程** ^[5]、**2011计划**重点建设的**全国重点大学**、**九校联盟** ^[7]、**中国大学校长联谊会**、**京港大学联盟** ^[8]、**亚洲大学联盟** ^[9]、**东亚研究型大学协会**、**国际研究型大学联盟**、**环太平洋大学联盟**、**东亚四大学论坛**、**国际公立大学论坛**、**中俄综合性大学联盟**重要成员。 ^[10-13]

北大始终与国家民族的命运紧密相连，聚集了许多学者专家，培养了众多优秀人才，创造了大批重大科学成果，影响和推动了中国近现代思想理论、科学技术、文化教育和社会发展的进程。 ^[14]

创办时间

1898年（**戊戌年**）7月3日

类别

公立大学

属性

九校联盟（C9）**985工程****211工程**2011计...

知名校友

李克强、**杨**振宁、**邓**稼先、**朱**自清、**屠**呦呦...

主管部门

中华人民共和国教育部

学校官网

<http://www.pku.edu.cn>

中文名	实体1	实体2
名称	北京大学	方正富邦基金管理 有限公司
别名	北大	方正富邦
所在省份	北京市	北京
统一社会信用代码		911100007178849 15E
工商注册资本		66000 万人民币
旗下基金数量		14

cnSchema2.0与图谱服务：KGAPI接口规范

KG第一境界：命名实体
更精准的意图理解

KG第二境界：实体属性模板
更精准的答案

实体分类	属性英文名	中文名	实体1	实体2
Thing	name	名称	北京大学	方正富邦基金管理 有限公司
Thing	alternateName	别名	北大	方正富邦
Organization	province	所在省份	北京市	北京
Company	unifiedSocialCredit Code	统一社会信用 代码		911100007178849 15E
Company	saicRegistrationCa pital	工商注册资本		66000 万人民币
MutualFundC ompany	numberOfFund	旗下基金数量		14

cnSchema2.0与图谱生产：实体属性模版

KG第二境界：实体属性模版（CnsTemplate）

标准名与概念定义

NLP

解析目标/溯源

版本质量校验

实体分类	属性	属性定义	属性中文名	原始属性值举例	预期属性类型	预期属性值举例	非空值覆盖率	值域一致性
Company	saicRegistrationCapital	注册资本是指合营企业在登记管理机构登记的资本总额,是合营各方已经缴纳的或合营者承诺一定要缴纳的出资额的总和。	工商注册资本	66000 万人民币	MonetaryAmount	{"value": 6600000000, "currency": "cny", "statedIn": "工商总局"}	95%	100%
Organization	province	组织机构所在的省份（直辖市以及特别行政区）	所在省份	北京市	Province	{ "@id": "...", "@type": ["Province", "Place", "Thing"], "name": "北京" }	100%	0%
MutualFundCompany	numberOfFunds	基金公司目前管理的基金总数	旗下基金数量	14	QuantitativeValue	{"value": 14, "endDate": "2018-08-01T00:00:00"}	90%	100%

cnSchema研发路线规划

	Schema	Tool
1.0 (17Q3)	• schemaorg3.2中文翻译与WIKI链接 • 命名实体 Thing	• cnSchema.org • github/cnSchema, 网站+概念搜索 • KGAPI, 知识图谱服务接口
2.0 (18Q3)	• 实体属性模版 CnsTemplate • schemaorg 3.4 增补更新 • 核心图谱 Schema • 实体二元关系 CnsLink	• KGTOOL, 图谱数据质量校验 • KGTOOL, Schema可视化
2.1 (18Q4)	• 本体计算 CnsOntology • 领域图谱Schema社区扩展	• KGAPI, 多层次图谱数据索引

cnSchema.org





OpenBase定位

OpenBase是OpenKG实现的开放知识图谱**众包**平台，主要目标是支撑**以中文为核心的开放领域知识图谱**的协同构建。

定位

促进中文文知识图谱**数据的开放与互联**，促进知识图谱和语义技术的普及和广泛应用

愿景

人人可参与，基于cnSchema，采用**CC0最彻底的开放许可协议**

特色

OpenBase需求分析

特定用户

知识图谱的使用者
(Bot, 搜索引擎等)

知识图谱创建者
(知识图谱工程师)

特定行业专家
(金融医药领域专家)

特定场景

了解数据, 使用图谱时

抽取, 编辑, 挖掘, 融合数据时

审核数据时

特定问题

构建高质量知识图谱的**时间成本**
图谱数据没有**可视化**的地方

构建过程实体链接**消歧问题**
schema的定义难
多数据源**融合**数据处理难
缺乏数据的**更新**机制

知识贡献**热情维系**
学习代价高, 友好程度不高

OpenBase流程图

基于NLP抽取+众包

知识的抽取与编辑

人工+自动化构建

cnSchema

扩展后

知识图谱数据来源，质量

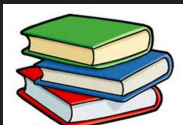
多数据源同一实体的不同定义
融合困难实体链接消歧

知识更新找不到方案，成本高

标注平台交互的友好程度不高
标注着缺乏热情
适合的审核员难以找到

OpenBase

原始数据



1.定义产品

2.基于cnSchema 做扩充

3.抽取和转化数据

1.多元的数据中挖掘相关的信息

2.融合相关数据

1.导入结构/非结构化的文本

2.通用或者专业领域的知识图谱

3.众包机制审核数据

抽取+编辑+挖掘+融合+更新

OpenBase

数据访问方式

SPARQL Endpoint

Linked Data

API

众包审核
存储

Graph DB

应用



Html Browser

1.定义知识更新的方式。

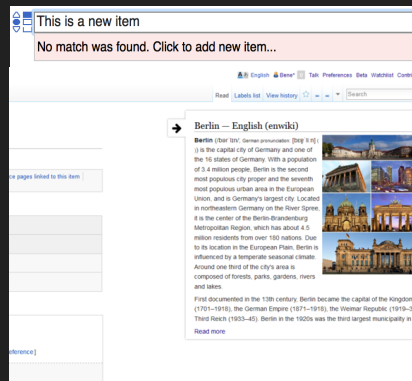
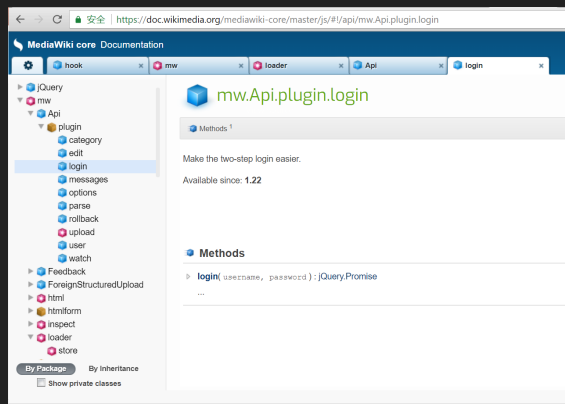
2.自动+人工的方式更新图谱数据

OpenBot

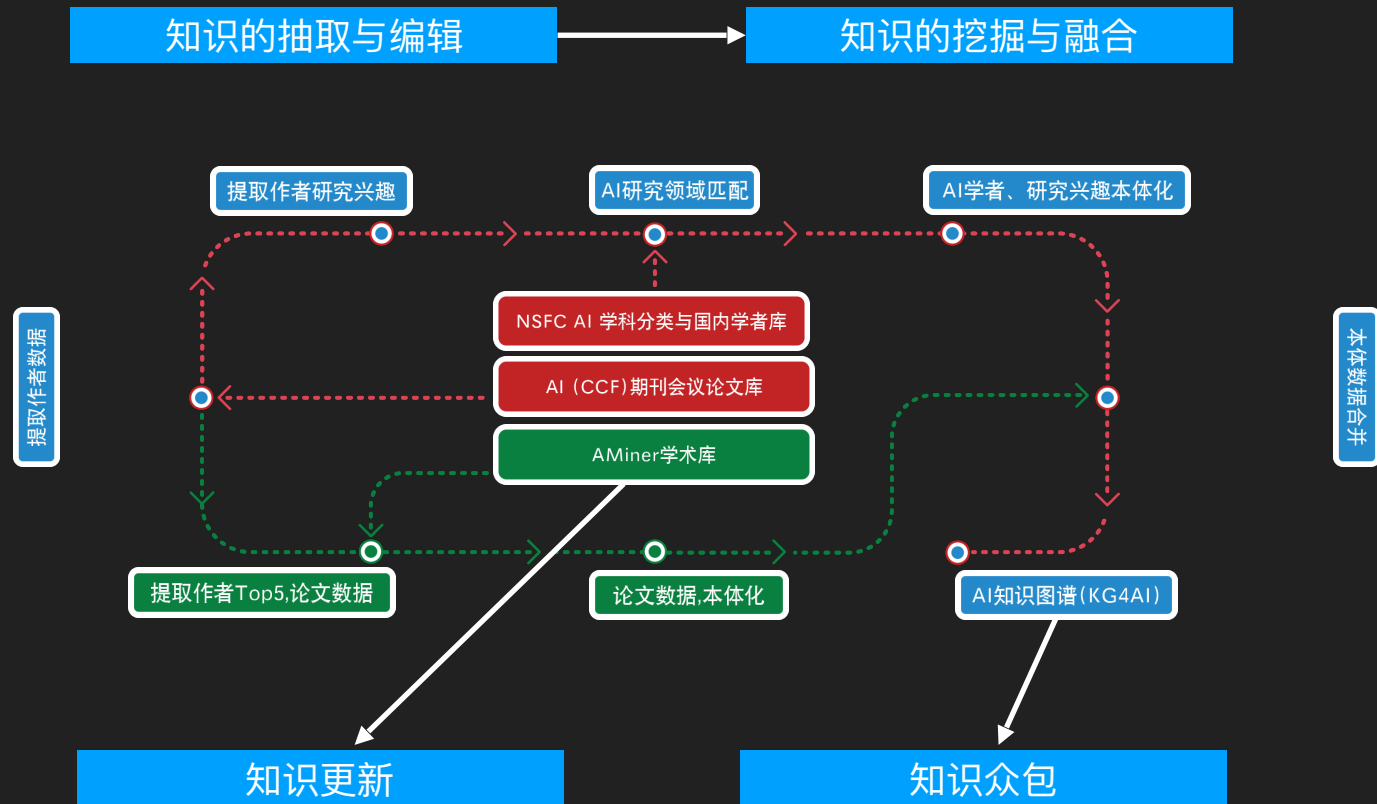
为什么没有基于WikiBase来做

Wiki Base限制太多

- 页面UI设计限制
- 插件功能限制
- 源码拓展较为困难
- 导入后台数据的效率问题



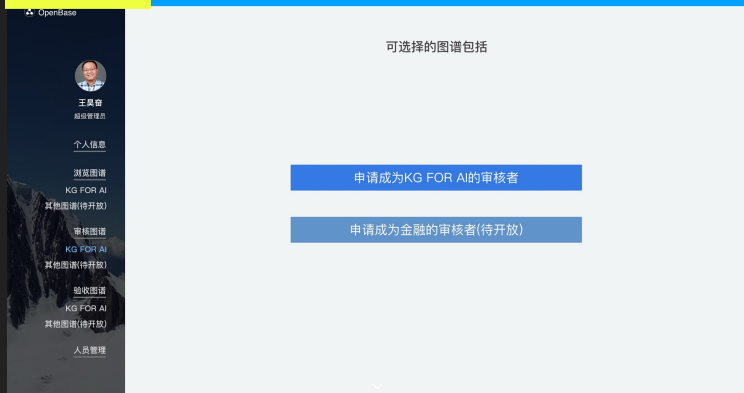
以KG4AI 全球人工智能学者知识图谱 为例



多角色联合

游客

按照专业能力选择图谱审核任务



游客

游客可以在测试通过后成为审核员



审核者

审核者可以在自己页面领取审核任务



验收者

验收者可以在自己页面领取验收任务



审核页面-实体维度

审核-实体维度

Openbase

王昊峰

超链管理

个人信息

浏览图谱

KG FOR AI

其他图谱(待升级)

审核图谱

KG FOR AI

其他图谱(待升级)

验收图谱

KG FOR AI

其他图谱(待升级)

人员管理

请挑选出下列不属于人工智能专家学者：

☒

刘欢

"刘欢 (Huan Liu)，现为美国亚利桑那州立大学 (Arizona State University) 计算机科学与工程的副教授。刘欢毕业于中国上海交通大学计算机科学与技术系计算机科学专业。"

☐

张长水

张长水，男，1985年生，河北人。智能技术系助理教授兼数据挖掘学术委员会委员。清华大学自动化系助理教授。主要研究领域为数据挖掘、机器学习、数据挖掘技术、数据挖掘应用、数据挖掘教育。发表学术论文100余篇。"

☐

梅宏

梅宏，男，1963年5月生，汉族，籍贯湖南，出生并成长于贵州。1984和1987年分别毕业于南京航空航天大学硕士和博士学位。1992年于上海交通大学获得博士学位。1994年在清华大学工作。现任清华大学教授。"

☐

刘欢

"刘欢 (Huan Liu)，现为美国亚利桑那州立大学 (Arizona State University) 计算机科学与工程的副教授。刘欢毕业于中国上海交通大学计算机科学与技术系计算机科学专业。"

☐

张长水

张长水，男，1985年生，河北人。智能技术系助理教授兼数据挖掘学术委员会委员。清华大学自动化系助理教授。主要研究领域为数据挖掘、机器学习、数据挖掘技术、数据挖掘应用、数据挖掘教育。发表学术论文100余篇。"

☐

梅宏

梅宏，男，1963年5月生，汉族，籍贯湖南，出生并成长于贵州。1984和1987年分别毕业于南京航空航天大学硕士和博士学位。1992年于上海交通大学获得博士学位。1994年在清华大学工作。现任清华大学教授。"

☐

黄庆明

黄庆明，男，博士，教授，国家杰出青年基金获得者。1984至1989年在哈尔滨工业大学计算机系攻读博士学位。1990年在香港科技大学计算机系工作。1990年在香港科技大学计算机系工作。主要研究方向为数据挖掘、数据挖掘应用、数据挖掘教育。发表学术论文100余篇。"

☐

胡事民

胡事民，清华大学计算机系副主任，学位分委员会主席，教授。1984年毕业于清华大学计算机系。1988年在清华大学计算机系工作。1990年在清华大学计算机系工作。主要研究方向为数据挖掘、数据挖掘应用、数据挖掘教育。发表学术论文100余篇。"

☐

戴琼海

戴琼海，1964年12月28日生，1990年博士毕业，清华大学自动化系教授、博导，智能系统研究所实验中心主任，主要研究方向为数据挖掘、数据挖掘应用、数据挖掘教育。发表学术论文100余篇。"

下一頁

验收-实体维度

Openbase

王昊峰

超链管理

个人信息

浏览图谱

KG FOR AI

其他图谱(待升级)

审核图谱

KG FOR AI

其他图谱(待升级)

验收图谱

KG FOR AI

其他图谱(待升级)

人员管理

请判断下列人员是否人工智能专家学者：

序号

实例1

关系/属性

审核结果

最终确认

1

刘欢

刘欢 (Huan Liu)，现为美国亚利桑那州立大学 (Arizona State University) 计算机科学与工程的副教授。刘欢毕业于中国上海交通大学计算机科学与技术专业。...

认为正确3人 认为错误2人

☒ 正确 ☐ 错误

2

张长水

张长水，男，1985年生，河北人。智能技术与数据挖掘系副教授兼数据挖掘学术委员会委员。清华大学自动化系助理教授。主要研究领域为数据挖掘、机器学习、数据挖掘技术、数据挖掘应用、数据挖掘教育。发表学术论文100余篇。...

认为正确12人 认为错误2人

☒ 正确 ☐ 错误

3

梅宏

梅宏，男，1963年5月生，汉族，籍贯湖南，出生并成长于贵州。1984和1987年分别毕业于南京航空航天大学硕士和博士学位。1992年于上海交通大学获得博士学位。1994年在清华大学工作。现任清华大学教授。...

认为正确34人 认为错误6人

☐ 正确 ☐ 错误

4

黄庆明

黄庆明，男，博士，教授，国家杰出青年基金获得者。1984至1989年在哈尔滨工业大学计算机系攻读博士学位。1990年在香港科技大学计算机系工作。1990年在香港科技大学计算机系工作。主要研究方向为数据挖掘、数据挖掘应用、数据挖掘教育。发表学术论文100余篇。...

认为正确32人 认为错误2人

☐ 正确 ☐ 错误

5

胡事民

胡事民，清华大学计算机系副主任，学位分委员会主席，教授。1984年毕业于清华大学计算机系。1988年在清华大学计算机系工作。1990年在清华大学计算机系工作。主要研究方向为数据挖掘、数据挖掘应用、数据挖掘教育。发表学术论文100余篇。...

认为正确3人 认为错误2人

☐ 正确 ☐ 错误

6

戴琼海

戴琼海，1964年12月28日生，1990年博士毕业，清华大学自动化系教授、博导，智能系统研究所实验中心主任，主要研究方向为数据挖掘、数据挖掘应用、数据挖掘教育。发表学术论文100余篇。...

认为正确3人 认为错误2人

☐ 正确 ☐ 错误

下一頁

审核-关系维度

Openbase

王昊峰

超链管理

个人信息

浏览图谱

KG FOR AI

其他图谱(待升级)

审核图谱

KG FOR AI

其他图谱(待升级)

验收图谱

KG FOR AI

其他图谱(待升级)

人员管理

正在审核:KG FOR AI

序号	实例1	关系/属性	实例2	操作
1	刘欢	新聘	教授	☒ 正确 ☐ 错误 关系不成立 ☐ 不知道
2	刘欢	李善材前		☒ 正确 ☐ 错误 关系不成立 ☐ 不知道
3	刘欢	李善材前	Huan Liu	☒ 正确 ☐ 错误 关系不成立 ☐ 不知道
4	刘欢	李善材前	113.8177	☒ 正确 ☐ 错误 关系不成立 ☐ 不知道
5	刘欢	李善材前	验收图谱	☒ 正确 ☐ 错误 关系不成立 ☐ 不知道

取消 提交

验收-关系维度

Openbase

王昊峰

超链管理

个人信息

浏览图谱

KG FOR AI

其他图谱(待升级)

审核图谱

KG FOR AI

其他图谱(待升级)

验收图谱

KG FOR AI

其他图谱(待升级)

人员管理

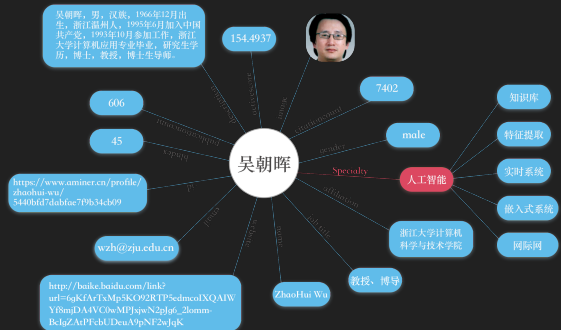
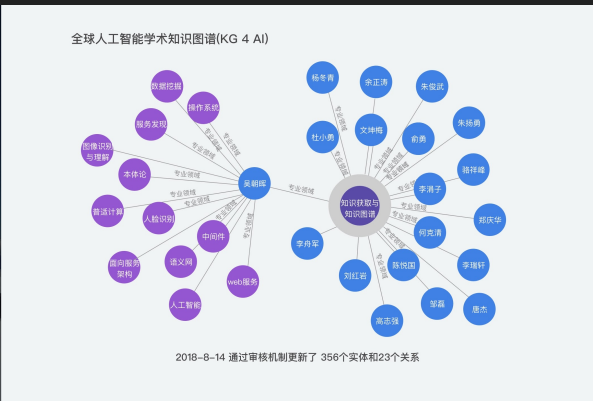
正在验收:KG FOR AI

序号	实例1	关系/属性	实例2	操作
1	刘欢	新聘	教授	认为正确3人 认为错误2人 详情 ☒ 正确 ☐ 错误
2	刘欢	李善材前		认为正确3人 认为错误2人 详情 ☒ 正确 ☐ 错误
3	刘欢	李善材前	Huan Liu	认为正确3人 认为错误2人 详情 ☒ 正确 ☐ 错误
4	刘欢	李善材前	113.8177	认为正确3人 认为错误2人 详情 ☒ 正确 ☐ 错误
5	刘欢	李善材前	验收图谱	认为正确3人 认为错误2人 详情 ☒ 正确 ☐ 错误

取消 提交

图谱的展示

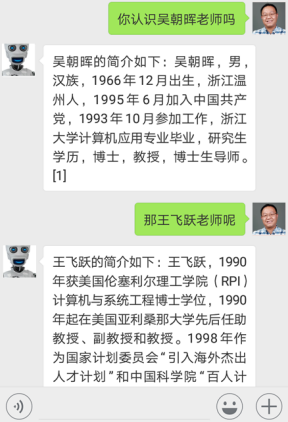
知识图谱的三种呈现方式： 图谱展示 对话Bot展示 实体视图展示



< KG4AI问答机器人



< KG4AI问答机器人



KG FOR AI

学者信息-吴朝晖

Name	ZHAOCHUI WU	Pubnumber	606
NameZH	吴朝晖	Email	wzh@zju.edu.cn
CitationCount	7402	Bio	吴朝晖, 男, 汉族, 1966年12月出生, 浙江温州人, 1995年6月加入中国共产党, 1993年10月参加工作, 浙江大学计算机应用专业毕业, 研究生学历, 博士, 浙江大学计算机科学与技术学院教授、博士生导师, 中国科学院院士。
Hindex	45	Website	http://baiko.baiko.com/link?url=6gKfArTsMpSkO92RTP5edmc0XQAIWYFlmDAAYC0wMPJspN2jg8_2lomm-B0zApF0bUDeuA9pN2wJgk
Affiliation	浙江大学计算机科学与技术学院		
Activity	154.4937		
Gender	male		
Job title	教授、博导		

社区运营-知识贡献榜

- 审核任务时间的排行榜；
- 审核任务的数量排行榜

加入Openbase社区

任务时间榜单

1 吴杨 6小时42分钟

2 李轩 5小时27分钟

3 朱慧源 5小时20分钟

4 李娟 5小时02分钟

5 姚光涛 4小时39分钟

用户贡献榜单

李轩 编辑321次

朱慧源 编辑290次

吴杨 编辑289次

姚光涛 编辑205次

李娟 编辑195次

立即加入

人员管理，个人信息管理

OpenBase

王昊奋

管理员

个人信息

浏览图谱

KG FOR AI

其他图谱(待开放)

审核图谱

KG FOR AI

其他图谱(待开放)

抽取图谱

KG FOR AI

其他图谱(待开放)

发送链接

人员管理

点击发送邀请链接

姓名	手机号码	邮箱	学校 / 机构 / 公司	擅长领域	注册时间	当前身份	操作
郑欢	12345678901	1234567890@qq.com	海知智能	人工智能	15小时	审核者	重新编辑
郑欢	12345678901	1234567890@qq.com	海知智能	知识图谱	34小时	已审核	重新编辑
刘欢	12345678901	1234567890@qq.com	海知智能	人工智能	15小时	审核者	审核者▼ 保存
刘欢	12345678901	1234567890@qq.com	海知智能	知识图谱	34小时	未审核	重新编辑
刘欢	12345678901	1234567890@qq.com	海知智能	人工智能	15小时	审核者	重新编辑
刘欢	12345678901	1234567890@qq.com	海知智能	知识图谱	34小时	未审核	重新编辑
刘欢	12345678901	1234567890@qq.com	海知智能	人工智能	15小时	审核者	重新编辑
刘欢	12345678901	1234567890@qq.com	海知智能	知识图谱	34小时	未审核	重新编辑
刘欢	12345678901	1234567890@qq.com	海知智能	人工智能	15小时	审核者	重新编辑
刘欢	12345678901	1234567890@qq.com	海知智能	人工智能	15小时	审核者	重新编辑

1 2 3 4 5 跳至 页

OpenBase

王昊奋

管理员

个人信息

浏览图谱

KG FOR AI

其他图谱(待开放)

抽取图谱

KG FOR AI

其他图谱(待开放)

发送链接

姓名: 王昊奋

手机号: 12345678901

邮箱: 12345678901@qq.com

学校 / 机构 / 公司: 深圳狗尾草科技

擅长领域: 人工智能

王昊奋的头像

编辑

OpenBase众包角色设计

	游客	审核员	验收员	超级管理员
目标人群	对OpenBase感兴趣的用户	行业从业者	特定领域专家	OpenBase人员管理团队
功能	知识图谱浏览	知识图谱浏览 知识图谱审核	知识图谱浏览 知识图谱审核 知识图谱验收	知识图谱浏览 知识图谱审核 知识图谱验收 人员管理
注册方式	开放式注册	1.开放制，游客答题通过后成为审核员 2.邀请制	邀请制	系统分配

项目kick off和推进花絮



★ openbase双周例会
23个参与者 | <https://join.skype.com/RIWcZeowdBf>

gtao y yao meng 陈华钧 丁力 帆玉
刘作鹏 鹏 张 漆桂林 孙胜男 王昊奋

20180710-众包平台-议程

会议目的: 明确众包项目目前
会议方式: skype: <https://join.skype.com/RIWcZeowdBf>
参会成员: 王昊奋、陈华钧、

名称 ↑
产品迭代
导入kg4ai
数据
众包平台
众包项目-尚忆秋-0720
openbase源文件与图片-Fina-2018.08.03
Openbase终版UI稿-2018.08.03

2018-08-08 openbase 周例会-议程
2018-08-11 openbase-纪要
2018-08-11 openbase-议程
20180403-众包平台-会议纪要.docx
20180416-周例会会议议程
20180416周例会会议纪要
20180416周例会会议流水账
20180502会议议程
20180502周例会会议纪要
20180509-schema会议纪要
20180509-schema会议议程
20180516-会议议程
20180516-周例会会议纪要

wikidata实现数据导入模块太慢-解决办法 #12 · opened a month ago by 陈阳
instance与schema关联 opened a month ago by wangxianmin Jul 10, 2018 Done
导入到wikibase opened a month ago by wangxianmin Done
e导入到wikibase ened a month ago by wangxianmin Jul 18, 2018 Done
e matching结果评估 ened a month ago by wangxianmin Jul 10, 2018 Done
kg4ai-数据统计-用于前端展示 #9 · opened 4 days ago by 陈阳
题库 #8 · opened a week ago by 陈阳
pi文档20180807 opened a week ago by 陈阳
/EBVOWL接口, 实现可视化编辑 opened a month ago by Feng.Gao Done
有代码如何与深交所众包项目结合, 是否有重用价值 opened a month ago by Feng.Gao Done
【众包 opened 2 months ago by 陈阳
【流程管理 opened 2 months ago by 陈阳 Done
【借鉴】实体抽取与标记 opened 2 months ago by 陈阳 Done

OpenBase参与成员



OpenBase网站开发组: 陈华钧、吴杨、李娟、顾进广、高峰、熊辉、姚光涛、杨帆王

OpenBase产品设计组: 谢殿侠、丁力、尚忆秋、王宇、李轩、朱慧源、杨帆、余华富、

KG4AI数据组: 李涓子、张鹏、唐杰、高博、张伟

项目管理与把控: 王昊奋、谢殿侠、王宇

OpenBase抽取与融合: 崔秋实、漆桂林、刘作鹏、李泰、王献敏、孙时、吴天星、李丞

KG可视化: 孟遥

cnSchema组: 丁力、陈华钧、王昊奋

法务组: 王宇

