开放科学框架及其实践

中国科学院文献情报中心
许哲平
xuzp@mail.las.ac.cn

虚拟天文台与天文信息学2019年学术年会 2019-11-27 黑龙江大庆

提纲

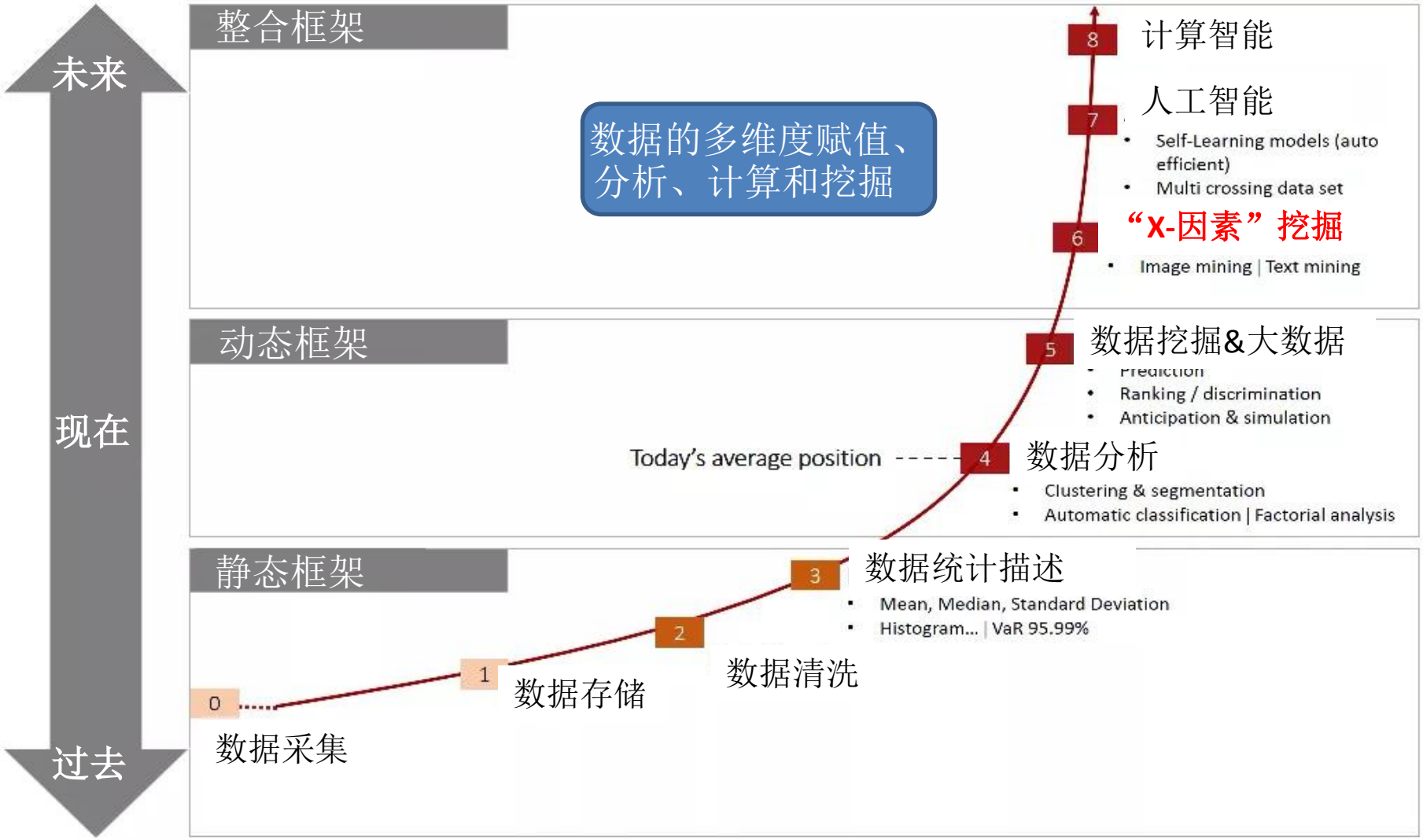
一、开放科学与数据科学

二、科研资源整合与关联

三、服务案例分析

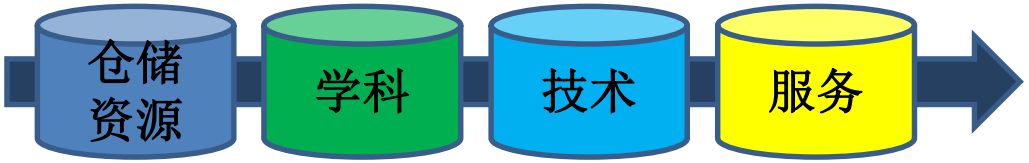
四、天文信息学

数据演化历史



Jeff Tuzik,2015：http://interfacejournal.com/archives/1407

大数据时代的科研维度



科研资源类型

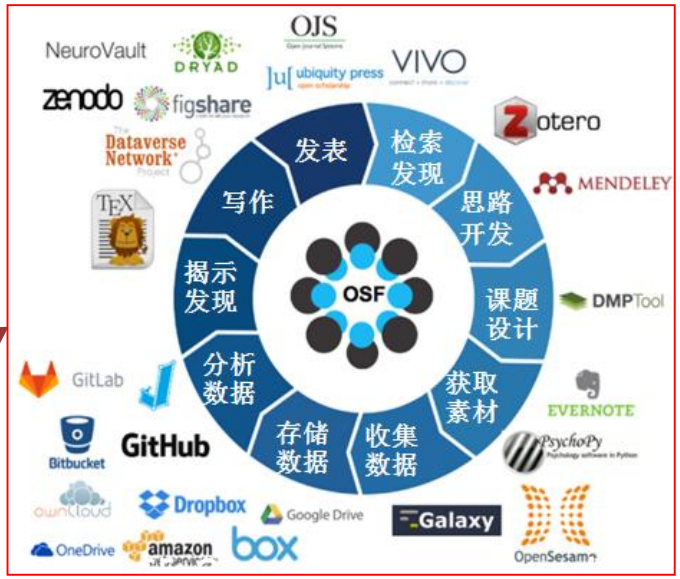
- 图书文献
- 科学数据
- 专利、报纸
- 传感网络
- GIS/遥感
- 社交媒体数据
- ……

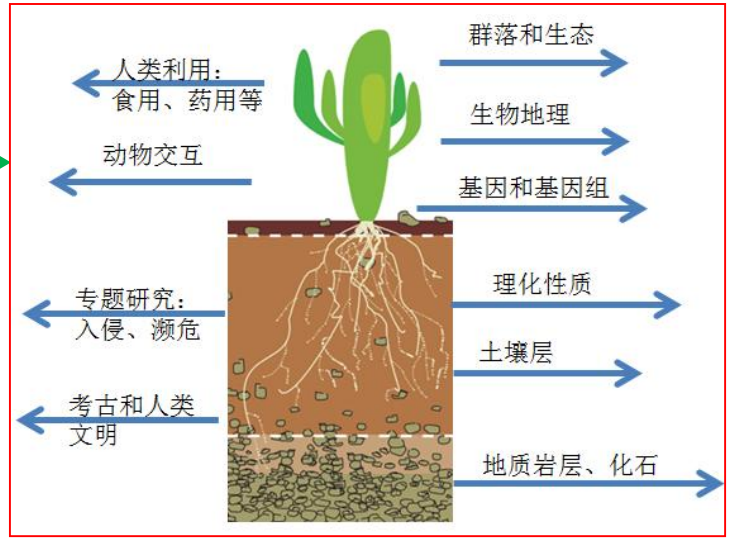
科研资源

我是谁？
我有什么？
我能干什么？

科研活动环节和流程

学科领域





通过自身积累以及对外服务和合作，积累科学数据资源的建设、整合、产品定制等方面的服务经验。紧跟Open Science Framework和GO FAIR等国际前沿理论和应用实践趋势。

数据科学（Data Science）
数据科学家（**Data Scientist**）

Astronomy vs Astroinformatics

Most of the initial time has been spent to find a common language among communities…

How astronomers see astroinformaticians

How astroinformaticians see astronomers

…with doubtful but promising results

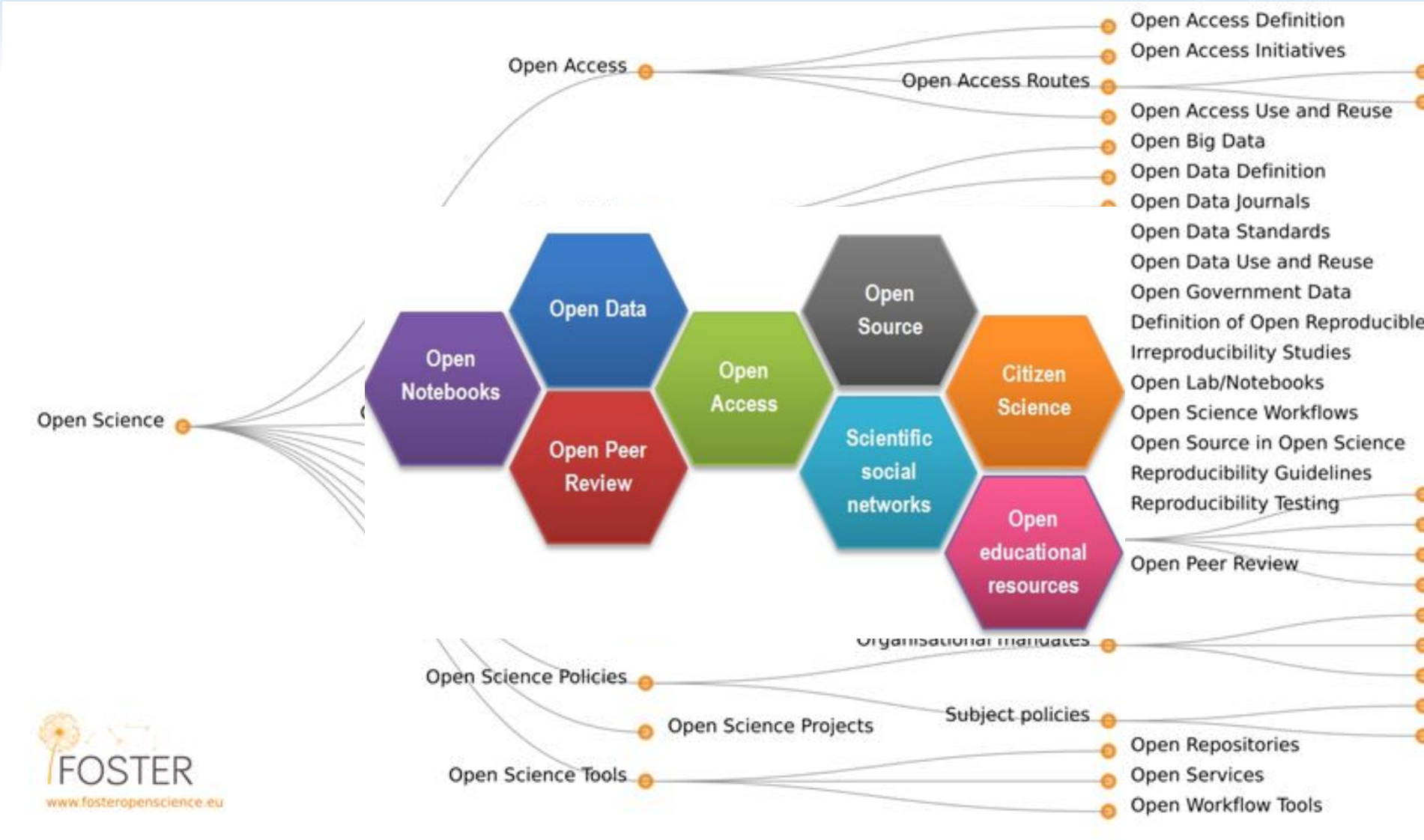
OSF（Open Science Framework）

开放科学框架

现状和背景

从开放获取(Open Access)==》开放研究（Open Science）

把各类资源当做整个科研基础设施的组成要素



提纲

一、开放科学与数据科学

二、科研资源整合与关联

三、服务案例分析

四、天文信息学

国外专业数据仓储

https://www.nature.com/sdata/policies/repositories

SCIENTIFIC DATA

View data repositories

- **Biological sciences:** Nucleic acid sequence; Protein sequence; Molecular & supramolecular structure; Neuroscience; Omics; Taxonomy & species diversity; Mathematical & modelling resources; Cytometry and Immunology; Imaging; Organism-focused resources
- **Health sciences**
- **Chemistry and Chemical biology**
- **Earth, Environmental and Space sciences:** Broad scope Earth & environmental sciences; Astronomy & planetary sciences; Biogeochemistry and Geochemistry; Climate sciences; Ecology; Geomagnetism & Palaeomagnetism; Ocean sciences; Solid Earth
- **Phy**
- **Mat**
- **Soc**
- **Gen**
- **Oth**

FAIR DATA
- Findable
- Accessible
- Interoperable
- Re-usable

FAIRsharing.org
standards, databases, policies

https://www.re3data.org/

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

https://www.re3data.org/

资源类型：拼图

资源 → 研究

- 物种2000中国节点
- The Plant List
- 物种2000全球节点
- International Plant Names Index (IPNI)
- 植物志
- 动物志

名录

- GBIF，全球最大
- NSII（国家标本资源共享平台）
- iDigBio，北美最大标本平台
- Atlas of Living Australia
- CForBio & SinoBON
- Global Plants on JSTOR
- **科技基础条件平台&科学数据库**

标本&观测

- CFH（中国自然标本馆
- PPBC（中国植物图像库）
- 中国鸟网
- 科学数据库中的生物图像
- EOL（Encyclopedia of Life）
- 澳大利亚植物图像索引
- 美国农业部植物图像集
- Wikispecies， wikipedia

图像

- BHL（生物多样性历史文献图书馆）
- BHL中国节点
- eFlora.org在线植物志项目
- 中国植物志在线版
- 中国动物志在线版

文献

- Barcode of Life Data Systems
- NCBI
- 中国西南野生生物种质资源库
- 中国珍稀濒危植物DNA条形码
- 中药材DNA条形码鉴定系统

DNA序列

- 澳门生物多样性数据中心
- 香港农业、渔业和保护部
- 台湾GBIF节点
- 四川省自然科技资源共享平台
- 广西植物资源信息共享平台

地方资源

“3S”数据资源

GPS：传感器、无人机、内置定位装置

GIS：GIS图层和GIS服务：WMS、TMS

RS：近地面遥感数据（如LIDAR数据等）

中国生物多样性监测网（完整性）

数据 → 指标 → 监测

数据的不完整性需要多方的共同努力和共享整合：We estimate that free-ranging domestic cats kill 1.4–3.7 billion birds and 6.9–20.7 billion mammals annually (*DOI: 10.1038/ncomms2380*)

北美地区每年被流浪猫致死的鸟达14-37亿只！

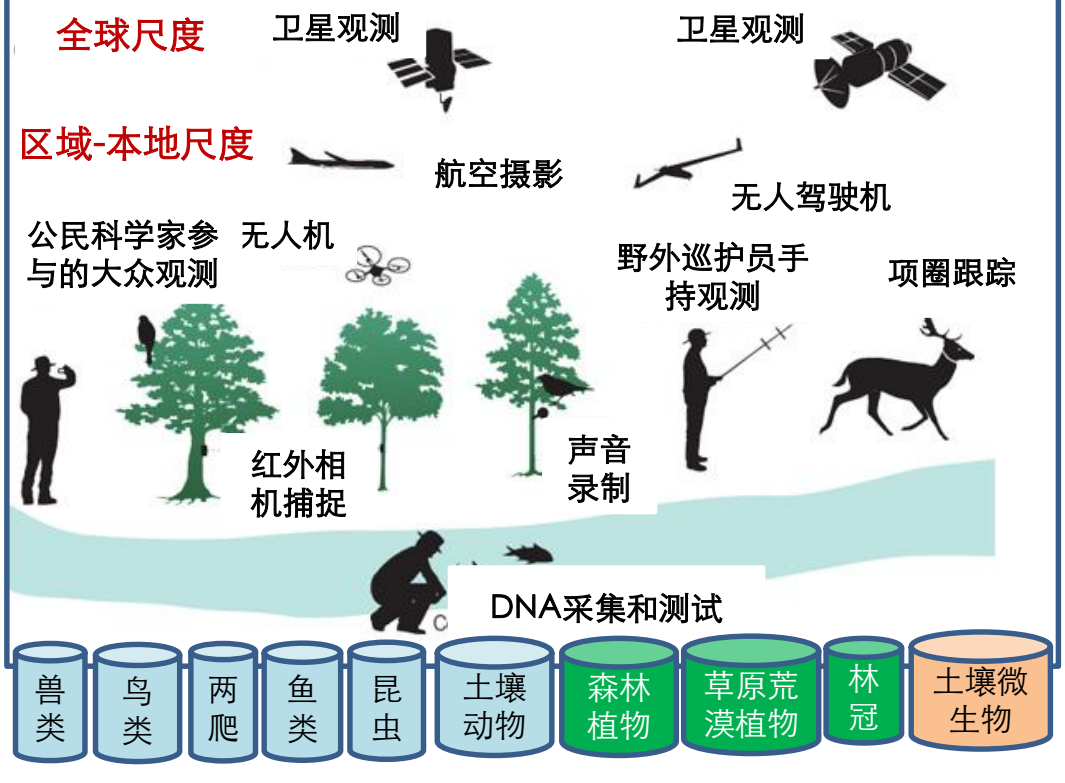
面上的无序堆积
VS
长期持续的定点监测？？

专业理论来做需求牵引和技术驱动：
Sino BON：中国生物多样性监测与研究网络（先导专项）

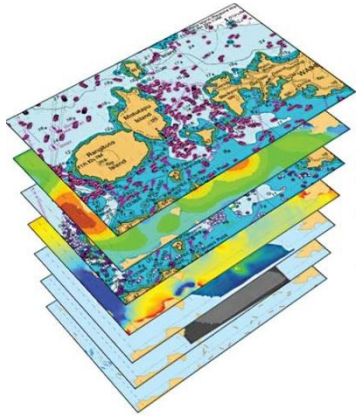
数据资源

中国生物多样性监测与研究网络(Sino BON)

植物多样性中心 | 动物多样性中心 | 微生物多样性中心 | 数据综合中心



挖掘分析



基于标准规范的各类生物多样性和生态图层

数据、工具和服务

外部数据整合 | 工具、模型、工作流

专题资源：
野生动植物栖息地保护
野生动物活动路线分析
典型地区3D植被和生境
气候变化对生境的影响
新技术方法的应用推广

服务

环保部： 生态安全监测预警及评估体系

林业局： 资源和生物多样性保护

农业部： 资源保护和生态修复

科技部： 国家科技重大专项、重大工程

国家十三五规划重大项目： 山水林田湖生态工程

国家十三五信息化规划： 国家大数据、美丽中国信息化、网上丝绸之路

服务类型及其扩展

- **资源类型的扩展（实例）**
- 文献：文章、图书、学位论文
- 专利
- 馆藏（标本/档案/图书/博物馆）
- 科学数据
- 多媒体
- 算法、代码
- **资源订购：数据服务**

- **知识元的扩展（系统）**
- 全国科学技术名词：20万+
- STKOS：60万+
- 中国分类主题词表：10万+
- Wikidata：500万+
- Ocid ID：500万+的作者，3000+万篇文章
- 各类领域本体

术语表（Term/Taxonomy）

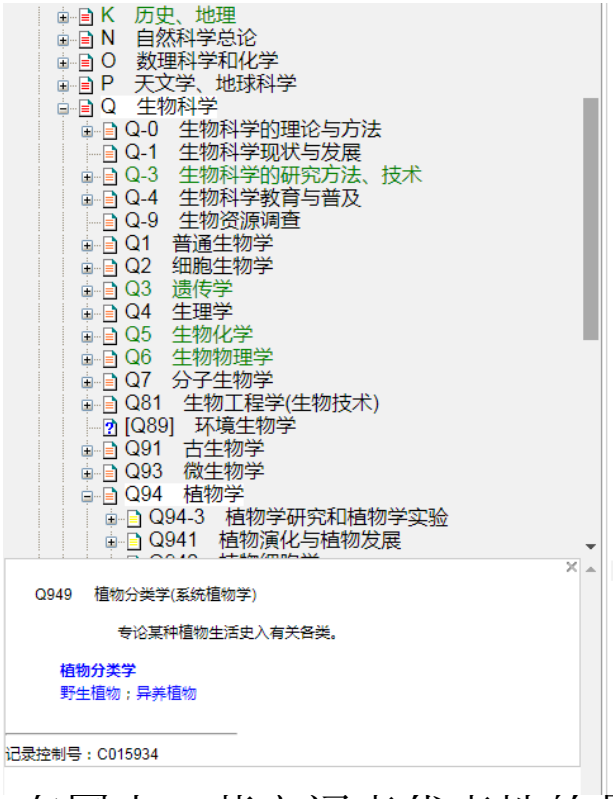
100多个学科，45万个术语
http://www.termonline.cn

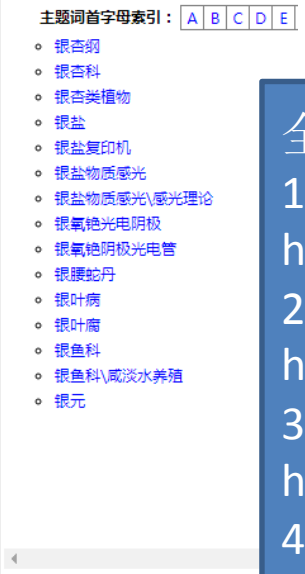
10+语言，关联数据映射
http://www.fao.org/faoterm/zh/

1.3万册工具书
http://gongjushu.cnki.net/

主题词表和知识组织体系（Thesaurus ）

中文：中国分类主题词表
http://cct.nlc.cn/login.aspx





全球农业主题词表
1.AGROVOC（多语种农业主题词表）
http://aims.fao.org/zh-hans/agrovoc
2.NALT（美国国家农业图书馆叙词表）
https://agclass.nal.usda.gov/
3.EUROVOC（欧盟农业主题词表）
https://eur-lex.europa.eu/browse/eurovoc.html
4.MeSH（医学主题词表）
https://www.ncbi.nlm.nih.gov/mesh/1000048

在国内，英文词表代表性的是科技知识组织体系(STKOS，http://stkos.las.ac.cn/）。该平台是建设以领域本体为目标的超级词表，包括来源术语、来源词表、科技术语、STKOS规范概念、范畴类和范畴表等6大类数据模型。目前平台收录61.5万学术概念，232.1万个术语，并对外提供了API、本体可视化和关联数据等多种功能。

元数据：扩展和映射

Astronomy

Metadata Standards

AVM - Astronomy Visualization Metadat
A standard defining discovery metada

FITS - Flexible Image Transport System
Used by the astronomy community to
including spatial, spectral and tempor

International Virtual Observatory Allian
A set of specifications, including meta
observatory.

SDAC - Standard for Documentation of
Used as an alternative to FITS for arc

SPASE Data Model
An information model for describing th

Extensions

FITS World Coordinate System (WCS)
An extension of FITS that enables dat
orignally proposed in 2002 then incor

IMPEx Data Model
A simulation extention to the SPASE

Resource Metadata for the Virtual Obse
Defines metadata terms and concepts

The extension is based on Dublin Cor

Tools

AVM Adobe Metadata Panels
A set of metadata panels that can be added to Adobe Creat
images.

AVM Web Tool
A web-based tool for assembling an AVM-compliant XMP

FITS Image Software Packages
Software packages that display or manipulate the relatively

GAVO DaCHS - Data Center Helper Suite
The software that underlies the German Astrophysical Virtu
Observatory-compliant data centres.

Saada
A tool for publishing astronomical data files as online databa

SDAC Tools
A set of four tools for working with SDAC-compliant archive
columns from a text file; anafile verifies that data files confo

SPASE Metadata Editor
A web-based editor for generating SPASE descriptions.

SPASE Tools
The SPASE website's list of tools for working with SPASE m

Use Cases

Aus-VO - Australian Virtual Observatory
An initiative to provide a distributed, uniform interface to the
astrophysical simulations, as part of the international Virtual

CDS - Centre de Données astronomiques de Strasbourg
CDS (Centre de Données astronomiques de Strasbourg/Str

http://www.dcc.ac.uk/resources/metadata-standards/disciplinary/astronomy

本体资源

- **OOR -** http://www.oor.net - index to the public instance(s) of the Open Ontology Repository initiative
- **OOR-sandbox -** http://sandbox.oor.net - OOR sandbox instance for shared open ontologies (at Northeastern University)
- **SOCoP-OOR -** http://socop.oor.net - the SOCoP-OOR on Ontohub instance by Spatial Ontology CoP (SOCoP) (at University of Magdeburg) **COLORE -** http://colore.oor.net - developing instance of the COmmon LOgic REpository (at University of Toronto)
- **Ontohub -** http://ontohub.oor.net - the developing Ontohub instance for distributed ontologies (at University of Bremen)
- **MMI-ORR -** http://mmisw.oor.net - the Ontology Registry & Repository for the Marine Metadata Interoperability (MMI) Project
- **BioPortal -** http://bioportal.bioontology.org/ - the open repository for Biomedical Ontologies by the National Center for Biomedical Ontology (NCBO)
- **Sigma-KEE -** http://sigma.oor.net - the open SIGMA Knowledge Engineering Environment for SUMO, MILO and related ontologies at OntologyPortal.org

Linked Data

http://linkeddata.org/

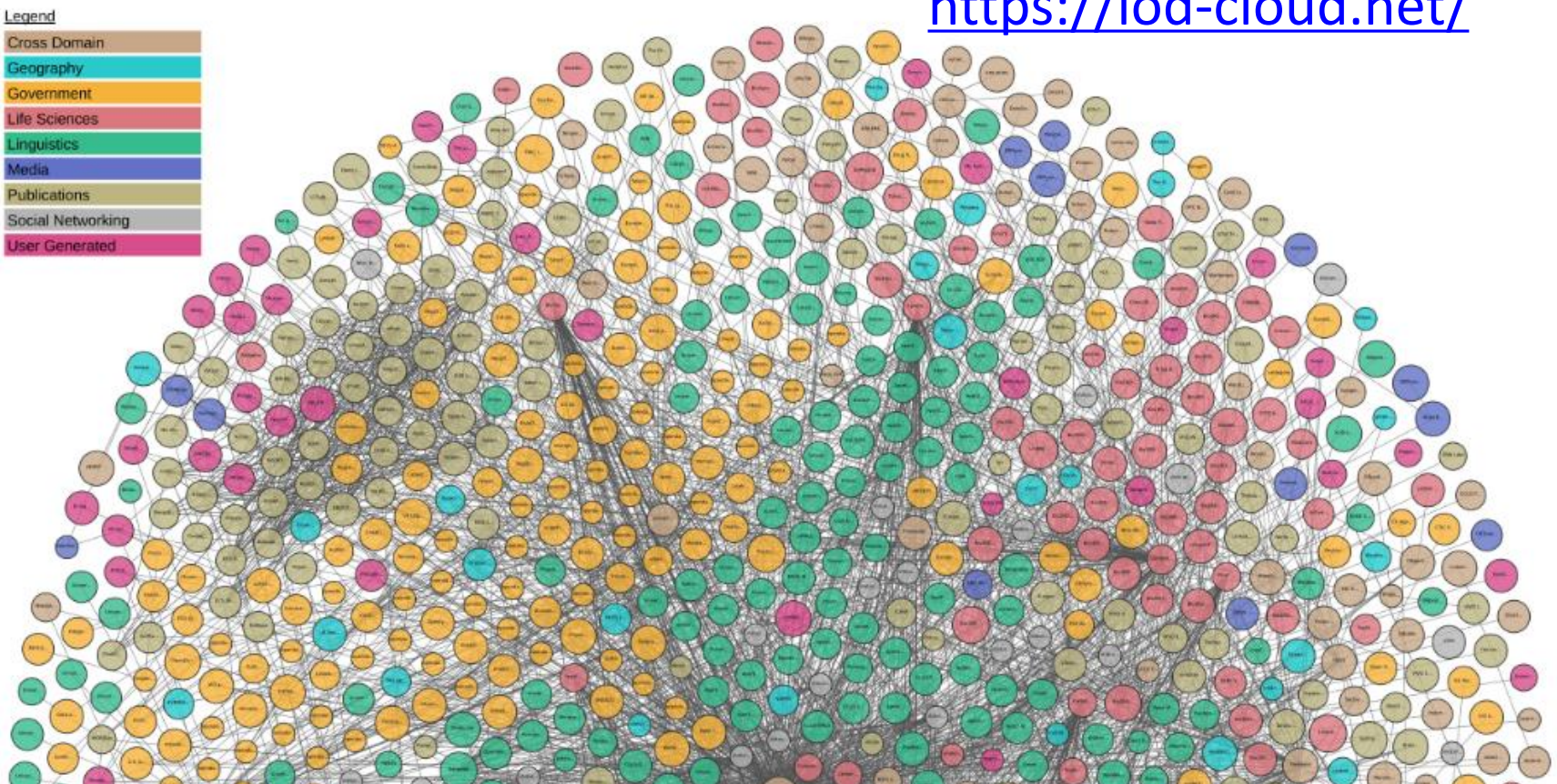
Linked Open Data，关联开放数据

关联数据是使用Web技术为无关数据建立相关性，或者使用Web技术来降低关联难度。更具体的是维基百科的定义——为了通过URI或者RDF技术来**暴露、共享和关联语义网上碎片化数据、信息和知识**，提供的最佳描述方案。

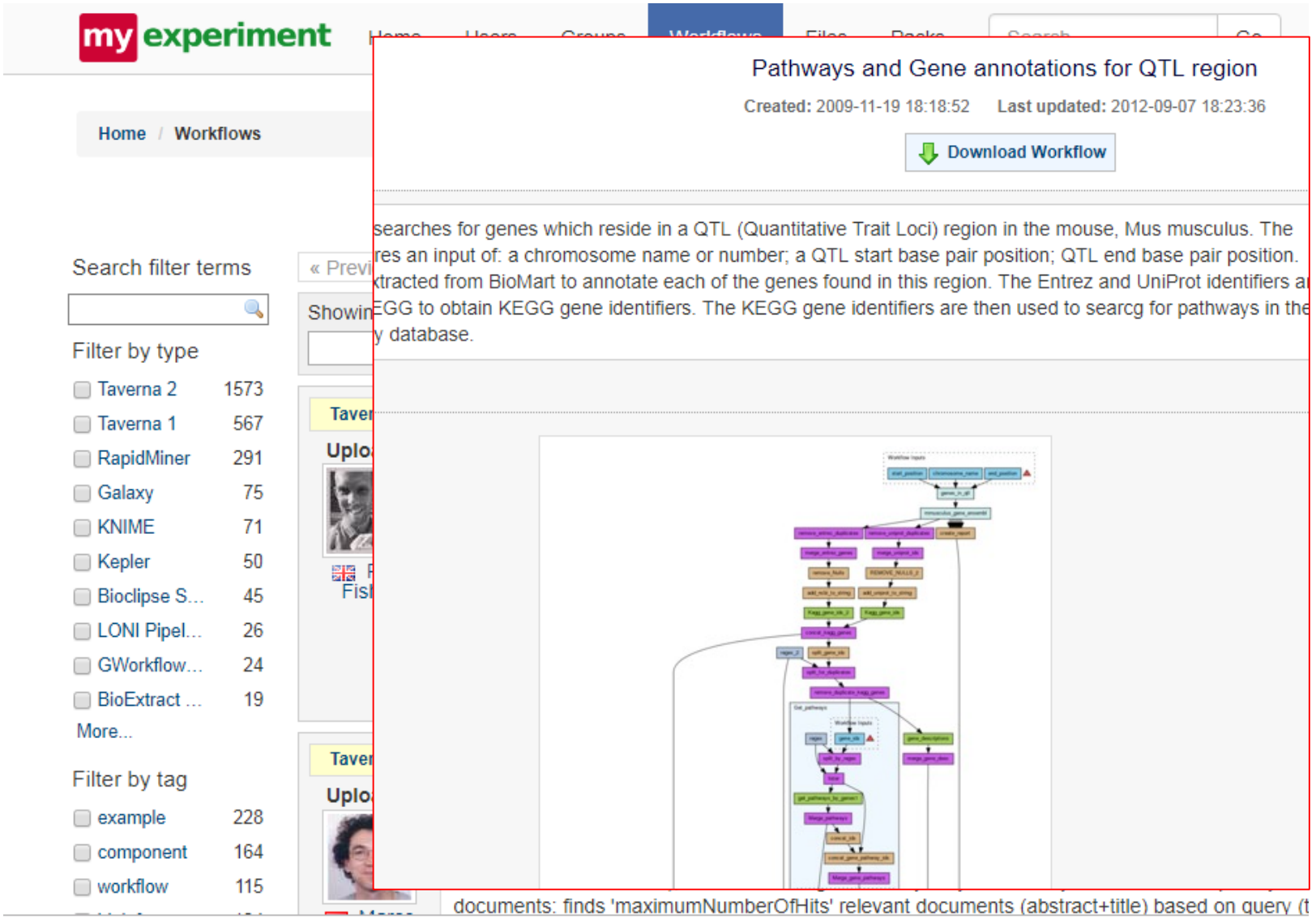
https://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets

The Linked Open Data Cloud

https://lod-cloud.net/

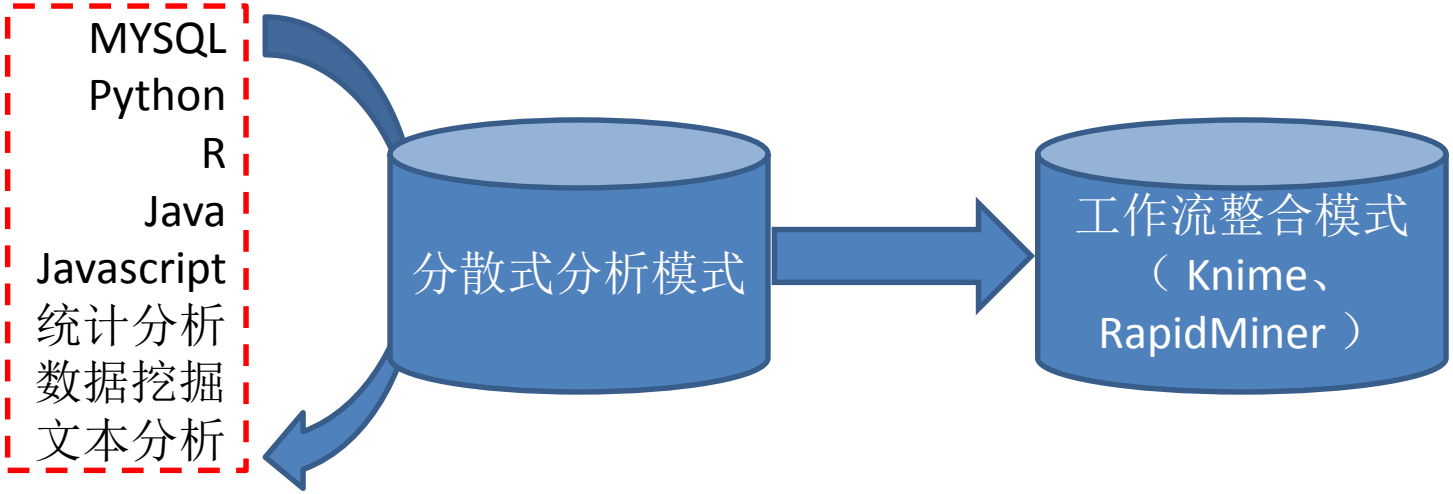


在线工作流分享：上传、共享、重用



数据挖掘工具

- 粗略分来，目前与数据挖掘及推荐引擎相关的开源项目主要有如下几类:
- **数据挖掘**相关:主要包括Weka、R-Project、Knime、RapidMiner、 等
- **文本挖掘**相关:主要包括OpenNLP、LingPipe、FreeLing、GATE 、Carrot2 等，具体可以参考LingPipe’s Competition
- **推荐引擎**相关:主要包括Apache Mahout、Duine framework、Singular Value Decomposition (SVD)，其他包可以参考Open Source Collaborative Filtering Written in Java
- **搜索引擎**相关:Lucene、Solr、Sphinx、Hibernate Search等



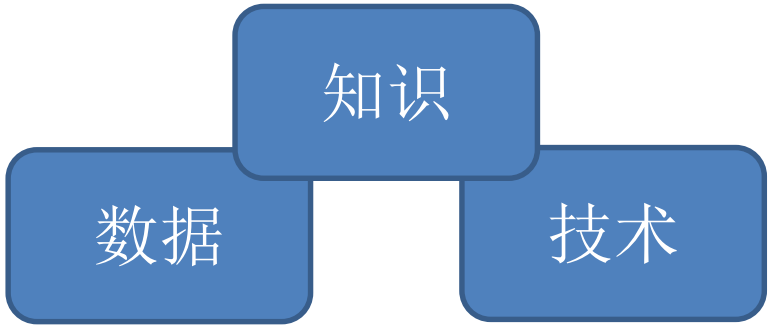
提纲

一、开放科学与数据科学

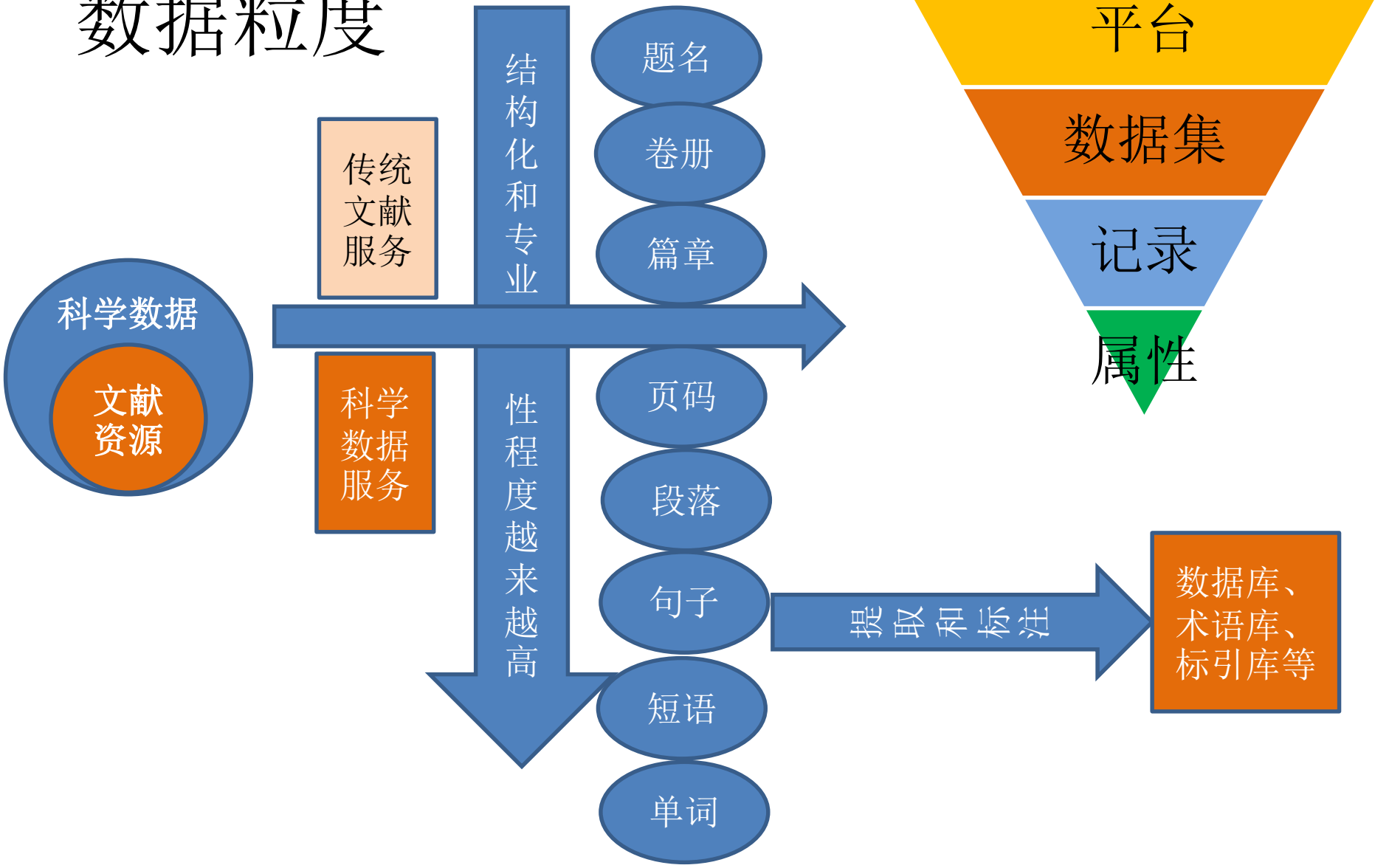
二、科研资源整合与关联

三、服务案例分析

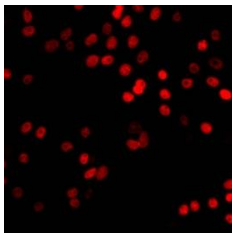
四、天文信息学



数据类型和数据粒度


科学数据
文献资源
传统文献服务
科学数据服务
结构化和专业性程度越来越高
题名
卷册
篇章
页码
段落
句子
短语
单词
提取和标注
数据库、术语库、标引库等
数据服务粒度
平台
数据集
记录
属性

服务案例：CV、NLP、ML



```
import cv2
import numpy
import sys
#Python + OpenCV : Score 27 Horizontal line scanni
#filename = sys.argv[1]
I = cv2.imread("50.jpg", 0)
h,w = I.shape[:2]
diff = (3,3,3)
mask = numpy.zeros((h+2
cv2.floodFill(I,mask,(0
T,I = cv2.threshold(I,1
I = cv2.medianBlur(I, 7

totalrice = 0
oldlinecount = 0
for y in range(0, h):
    oldc = 0
    linecount = 0
    start = 0
    for x in range(0, w
        c = I[y,x] < 2
        if c == 1 and o
            start = x
        if c == 0 and o
            linecount +
        oldc = c
    if

print(
print(totalrice)
```

```
管理员: 命令提示符
Microsoft Windows [版本
版权所有 (c) 2009 Micro

C:\Users\Administrator

C:\Users\Administrator
米粒数量是:
50

C:\Users\Administrator
```

CV：计算机视觉

重要病虫害数据库

选择图片： 选择文件 未选择任何文件 提交



ML：机器学习与图像识别

地名 人名 时代（年份） 著作 专有名词 引用注释 动植物名称

浙江省农业志（上册）

一、蚕业溯源

浙江是全国蚕桑生产发源地之一。考古发掘表明，9500～7500年前就有桑树出现。在余姚河姆渡新石器时期遗址发掘中，出土有结果相子、果核、梭形器、木制纺纱棒、打纬刀、经轴、陶制纺轮等纺织工具和器外口径4.8厘米、高2.4厘米。蚕纹形象生动，河姆渡人已认识蚕茧的利用价值，茧丝纤维尚未一致定论。对野蚕是否已进入家养阶段，织原料。吴兴钱山漾遗址出土的绸片、丝带和，绸片长2.4厘米，宽1厘米，呈黄褐色，没缫制而成，没有加捻，丝缕平直，呈平纹组织径167微米，组成绸片茧丝呈现半透明状，镜一团，有些片段上残留着绸片。丝线呈Z捻，典型的桑蚕丝结构。这是世界上迄今发现最早史。

《越绝书》以春秋末年至战国初期吴越争霸的历史事实为主干，上溯夏禹，下迄两汉，旁及诸侯列国，对这一历史时期吴越地区的汉民族政治、经济、军事、天文、地理、历法、语言等多有所涉及，被誉为"地方志鼻祖"。

古籍善本(国家级重点古籍名录)：

003 越绝书十五卷（汉）袁康撰 明刻本 孙诒让 浙江大学图书馆

0 越绝书十五卷（汉）袁康撰明万历十四年 586）冯念祖卧龙山房刻本 陕西省图书馆

州）产有"贡锦"，即以贡国被吴国打败。越王句践在（卷第四）记载，句践得范蠡之"，发展蚕桑生产。当时丝织品

NLP：自然语言处理与批量标引、分词、新词、分类和NER

PLOS ONE

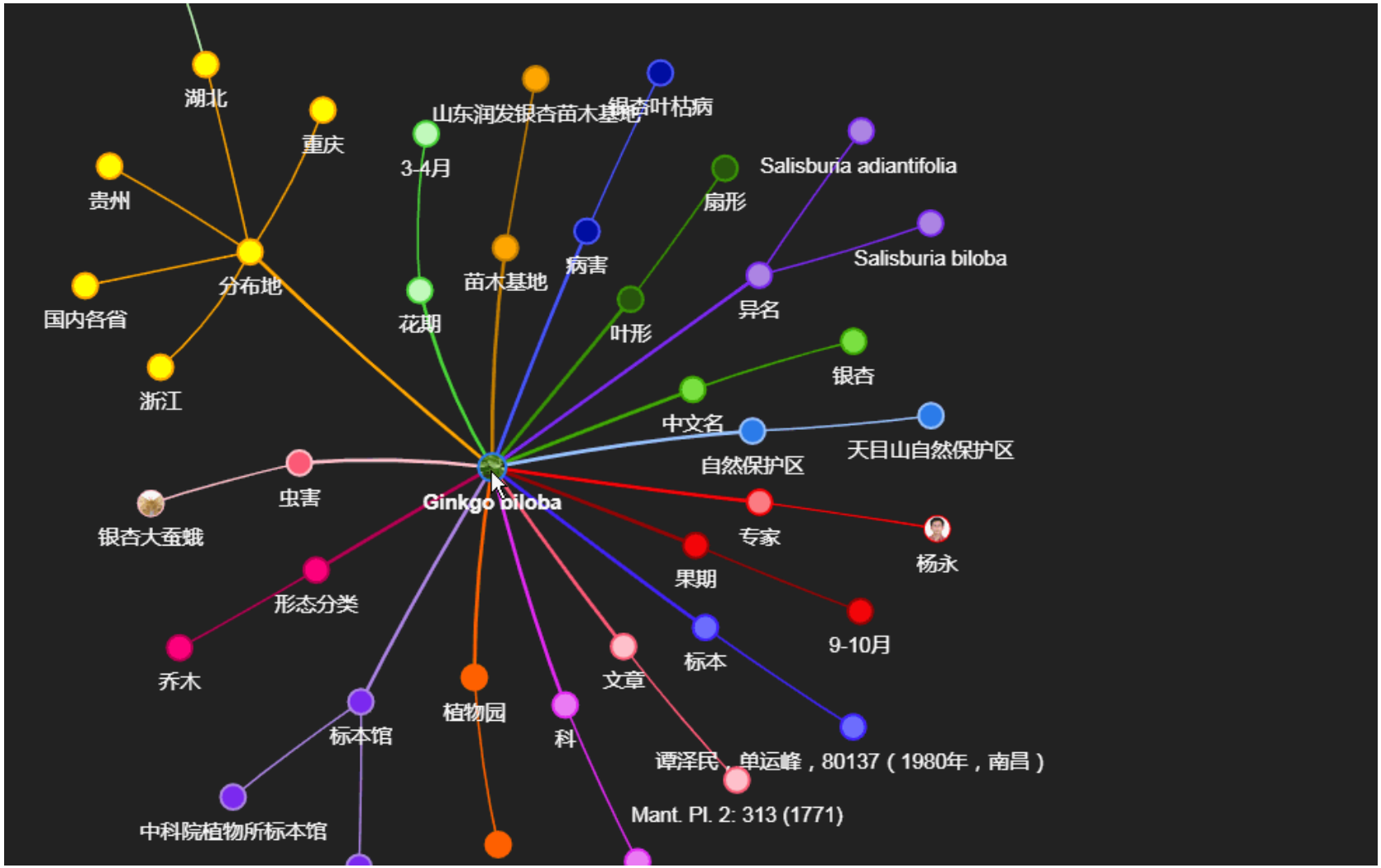
Sequencing data has been submitted to the National Center for Biotechnology Information (NCBI) under BioProject ID **PRJNA208899**.

NCBI Resources

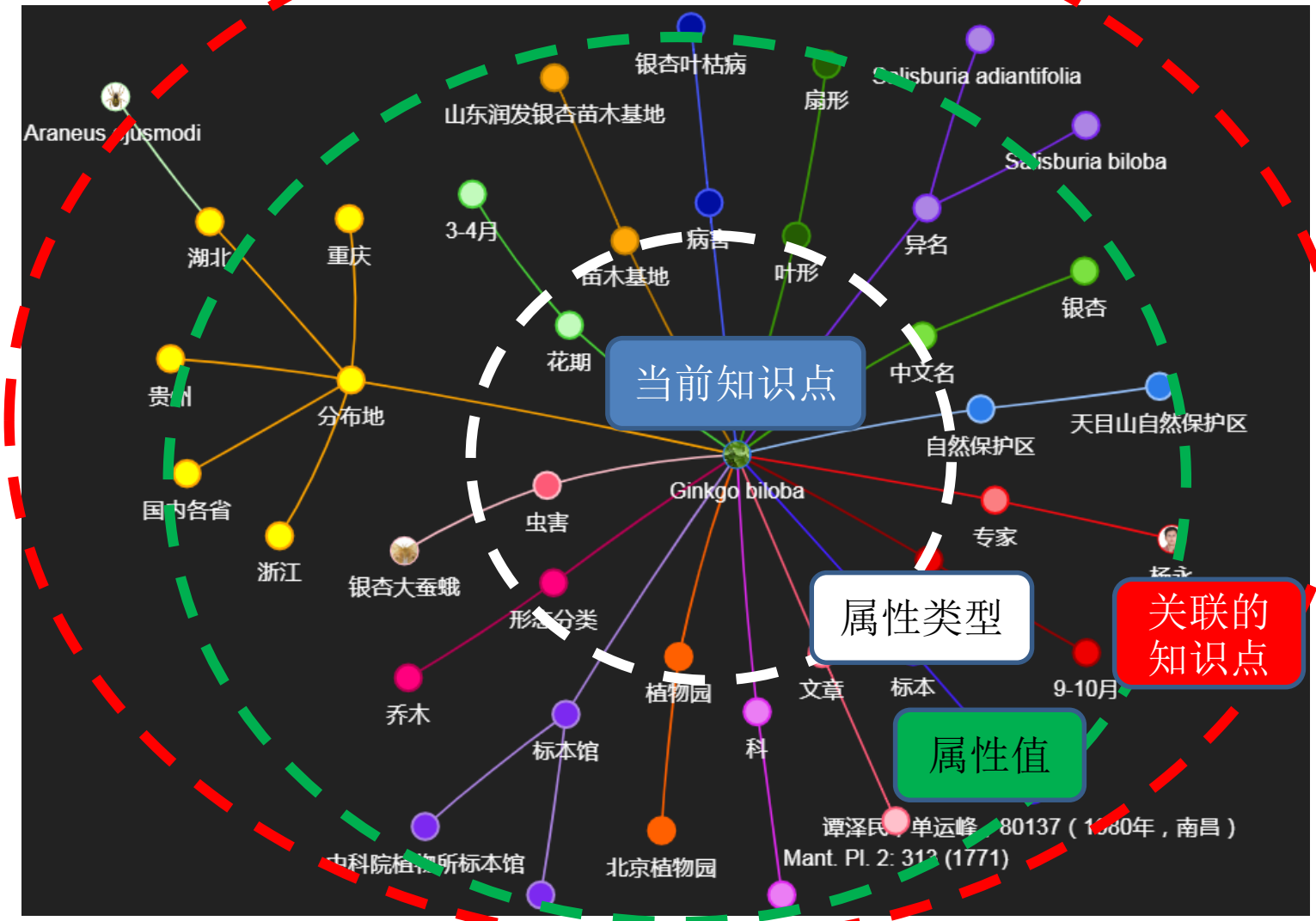
BioProject

新一代地方志：多源数据整合的增强型志书系统

数据标注与知识图谱

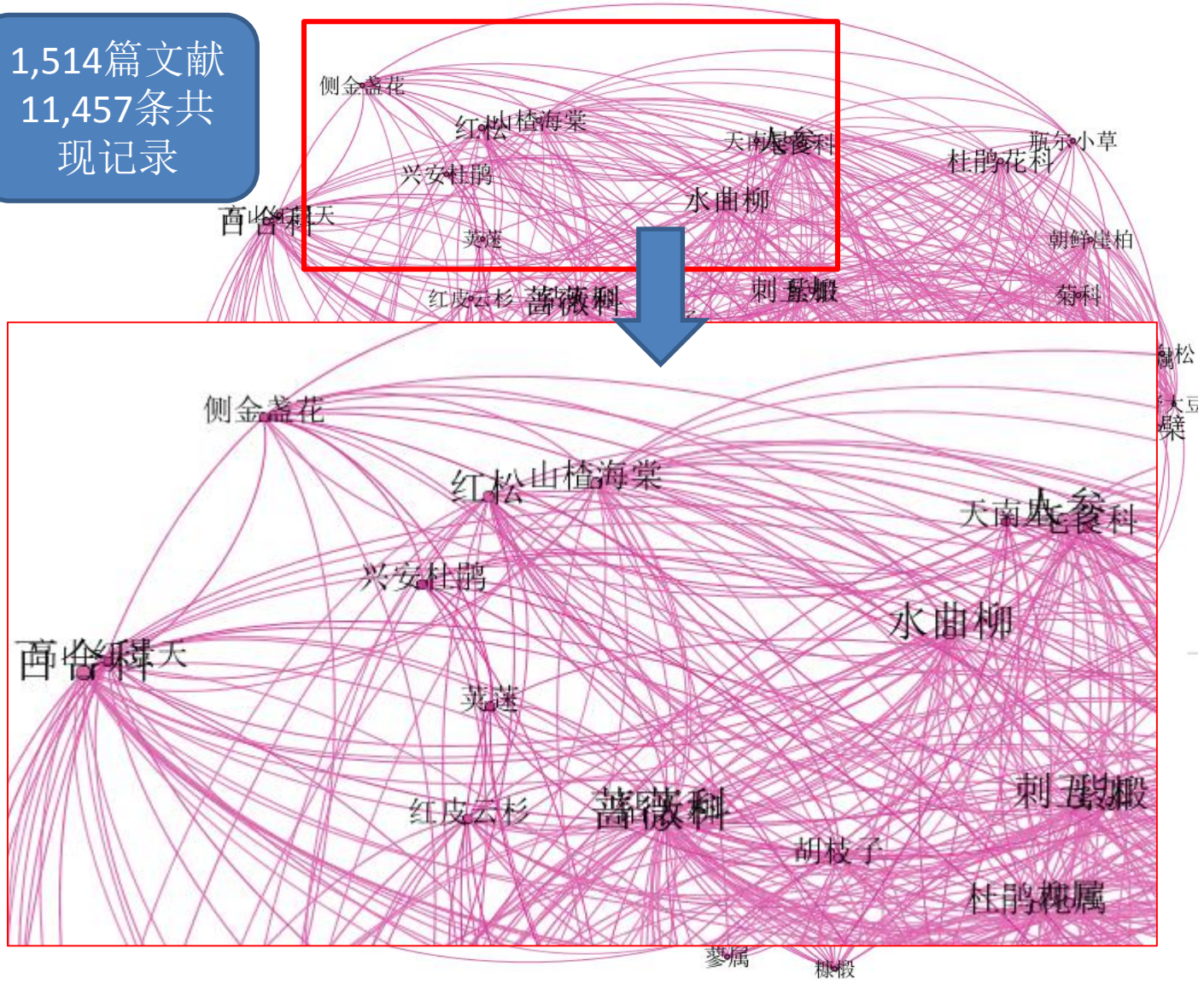

湖北
重庆
贵州
分布地
国内各省
浙江
山东润发银杏苗木基地
银杏叶枯病
3-4月
苗木基地
病害
花期
叶形
扇形
Salisburia adiantifolia
Salisburia biloba
异名
银杏
中文名
天目山自然保护区
自然保护区
虫害
Ginkgo biloba
银杏大蚕蛾
专家
杨永
果期
9-10月
形态分类
乔木
标本
文章
植物园
科
标本馆
谭泽民，单运峰，80137（1980年，南昌）
Mant. Pl. 2: 313 (1771)
中科院植物所标本馆

知识图谱与圈层结构分析



长白山地区的植物共现关系

1,514篇文献
11,457条共现记录

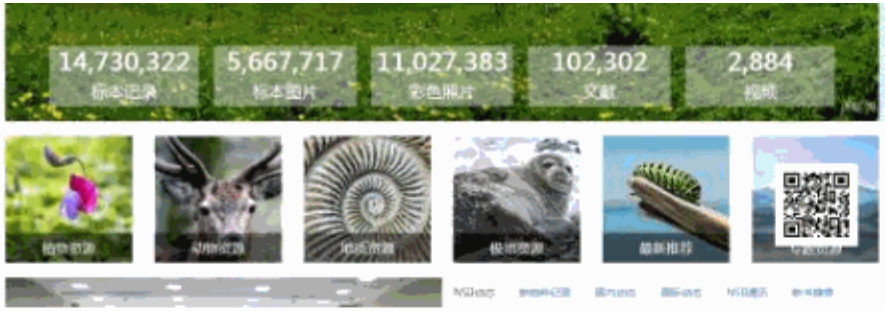


物种1	物种2	共现数量
水曲柳	红松	79
紫椴	红松	74
红松	蒙古栎	58
红松	色木槭	53
紫椴	蒙古栎	48
白桦	红松	46
水曲柳	紫椴	45
冷杉	红松	44
紫椴	色木槭	43
水曲柳	色木槭	40
水曲柳	蒙古栎	38
椴树	红松	38
色木槭	蒙古栎	37
山杨	白桦	36
云杉	冷杉	35
山杨	红松	34
岳桦	红松	32
云杉	红松	32
水曲柳	白桦	30
红松	落叶松	29
春榆	水曲柳	29
山杨	水曲柳	27
春榆	红松	25
春榆	色木槭	23
春榆	紫椴	23
山杨	紫椴	22
春榆	蒙古栎	21
白桦	色木槭	21
白桦	紫椴	21

中国科学数据仓储系统登记和共享平台

Registry and Sharing Platform of China Research Data Repository

首页 仓储推荐 仓储列表 仓储地图 关于我们 注册登录



（2005DKA21402）、教学标本（2005DKA21403）、保护区标本（2005DKA21404）、岩矿化石标本（2005DKA21405）和极地标本（2005DKA21406）6个子平台，截止2018年底有198个参加单位组成，涉及到中国科学院、教育部、国土资源部、国家海洋局和国家林业局等主管部门。平台目前工作人员有285人，其中运行管理人员52人、技术支撑人员153人、共享服务人员80人，共同完成平台的运行和服务。

关键字: 生物多样性 植物

省份/州:

北京

学科分类:

Q91-古生物学, Q93-微生物学, Q94-植物学, Q95-动物学

牵头建设单位:

中科院植物所

其他参加单位:

中科研动物所, 中科院昆明植物所, 中国林科院, 中国地质大学（北京）

相关链接:

首页

http://www.nsii.org.cn

资源类型:

图书, 数据集, 期刊文章

仓储语言:

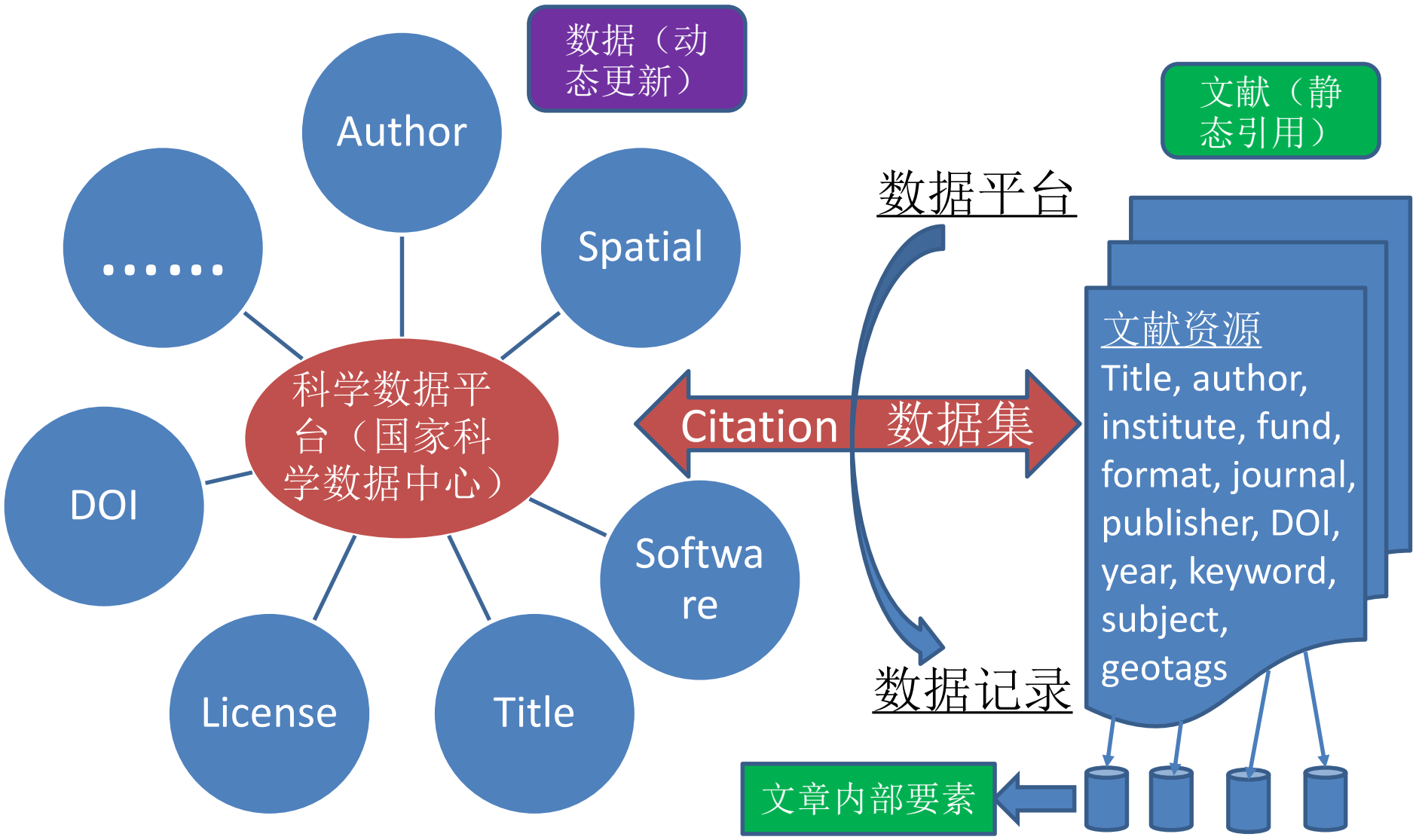
中文, 英语

国家:

中国

地图位置

科学数据与文献资源的关联和交互



数据引用和文献关联专题资源库

首页 | 数据引用专题库

25个国家科学数据平台的2610条引用信息

项目介绍

本项目主要收集和整理科学数据引用的相关文献记录，建立两者之间的关联，前期主要收集和整理科技部基础条件平台支持的28个国家级科学数据平台上的数据引用情况，并对文献进行分类整理，形成专题资源数据库。

科学数据平台导航

- 国家地震科学数据共享服务平台
- 中国数字科技馆
- 国家科技图书文献共享服务平台
- 国家气象科学数据共享服务平台
- 国家基础科学数据共享服务平台
- 国家地球系统科学数据共享服务平台
- 国家农作物种质资源共享服务平台
- 中国生态系统研究网络
- 国家材料环境腐蚀野外科学观测研究共享服务平台
- 国家人口与健康科学数据共享服务平台
- 国家农业科学数据共享服务平台
- 国家标本资源共享服务平台
- 国家林业科学数据共享服务平台
- 国家人类遗传资源共享服务平台
- 国家重要野生植物种质资源共享服务平台
- 国家标准文献共享服务平台
- 国家生态系统观测研究共享服务平台
- 国家微生物资源共享服务平台
- 国家标准物质资源共享平台
- 国家水产种质资源共享服务平台
- 国家实验细胞资源共享服务平台
- 国家计量基标准资源共享服务平台
- 国家家养动物种质资源共享服务平台
- 国家应急分析测试共享服务平台

清代河南赋税数据库建立方法研究[J]

匿名 (未验证) 在 周五, 08/09/2019 - 09:04 提交

平台名称: 国家地球系统科学数据共享服务平台

查看更多 登录或注册后发表评论

NSTL科技热点门户与重点领域网络信息跟踪服务系统的比较分析[J]

匿名 (未验证) 在 周五, 08/09/2019 - 09:04 提交

平台名称: 国家科技图书文献共享服务平台

或注册后发表评论

亚热带天然次生混交林生物量及养分生物循环研究[D]

匿名 (未验证) 在 周五, 08/09/2019 - 09:02 提交

平台名称: 国家生态系统观测研究共享服务平台

来源:
中南林业科技大学

年卷期页:
2011.

文章ID:
3E.A.E.U.M.F.6R.U3I.V.MJV.QF.

数据ID:
1012258761

数据引用:
[159]国家生态系统观测网络2004.http://www.cnern.org/web/index3.aspx?menu ID=2292

或注册后发表评论

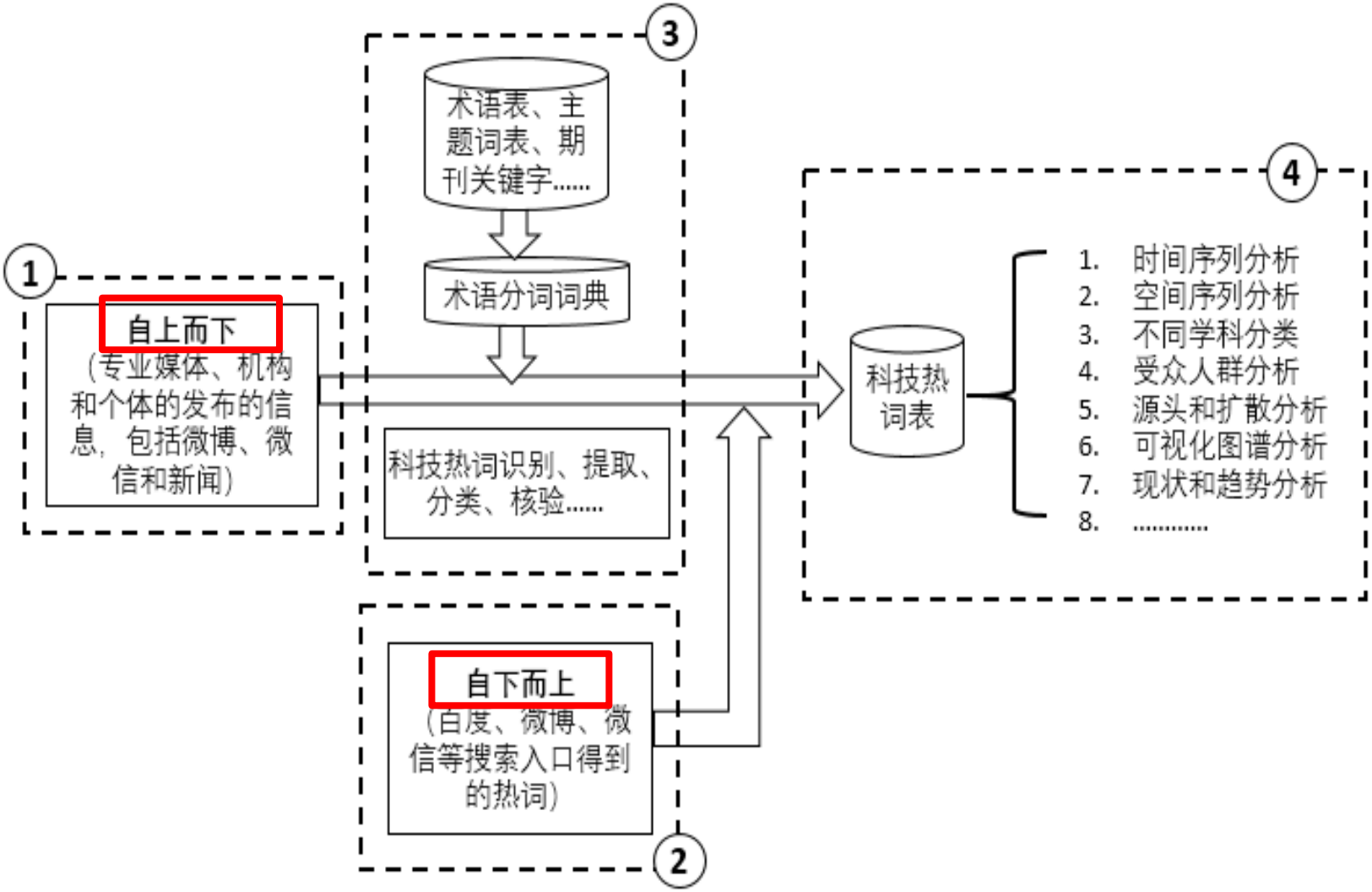
或注册后发表评论

数据引用-文献关联结构

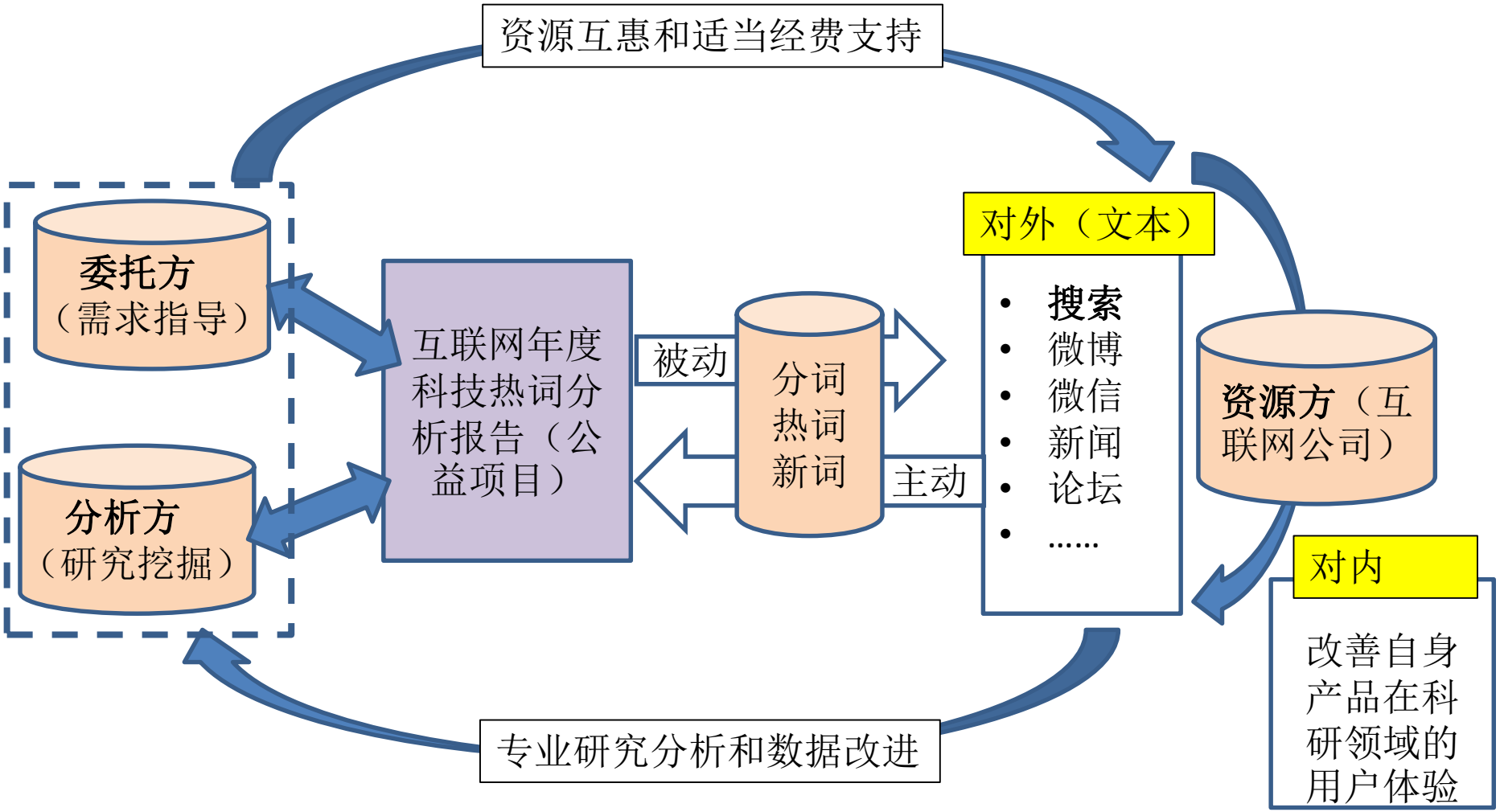
分类					
科学数据部分	文献引用部分	引用位	作者		
			单位		
		正文参	年份		
			来源（期刊名学位学校和专出版社）		
		参考文	卷期页		
			分类号		
			关键字（多条录用逗号分隔		
		其他位容（脚			
		引用文（doi）	摘要		
		文章U或万方	基金（多条记用@@分隔）		
		文章U官网）			
		标题			

专业术语标引		皱蒴藓属(Aulacomnium)和赛藓属(Meesia)	
研究地区	芦山地震主震震中及其周边地区	中国	
研究时间	2013 年 4 月 26-2014 年 5 月 8 日		
试剂（多条记录用@@分隔）			1.1.2 培养基 LB 培养基：葡萄糖 5 g/L，蛋白胨 10g/L,酵母膏 5g/U NaClOg/L, pH7.0; YPG 培养基：酵母提取物 5 g/L。
仪器（多条记录用@@分隔）	SercelL 22E 短周期探头		1.1.4 主要设备 PCR 仪：德国 BIOMETRA 公司； 高压脉冲电击转化仪、全 CI 动凝胶成像仪、等电；
装置（多条记录用@@分隔）			
方法（多条记录用@@分隔）	该程序使用了 Geiger（1912）提出的经典思想，即将非线性方程组线性化，通过最小二乘原理求解．	接受者操作特性曲线（Receiver operating characteristic，ROC）	1.2.1 DNA 操作方法 Ecoli 质粒提取、Ecoli 感受态 细胞制备与 CaClg2。AcHobater pasteuriamm 基因组的提取参照文献。
模型（多条记录用@@分隔）	一维波速模型的改进：赵珠等（1997），Wang 等（2007）	Maxent 模型	
软件工具（多条记录用@@分	Hypo71	Maxent 3.3.3k 软件 @@ArcGIS 10.2 空间分析	

互联网年度科技热词舆情分析报告预研



互联网年度科技热词舆情分析报告预研



APP和微信公众号开发

公众号进入后点击
“资源中心”即可

知识服务
微信号: pavoinfo

教育与公民科学

中科院机构知识库的服务案例：中学素质教育扩展（北师大）

树木年轮

反映了太阳的活动规律

太阳耀斑会放出碳14，树木光合作用吸收以后进入木质部，保存在年轮；

对地震研究的启示

地震的震动会造成年轮偏窄，研究这些规律从而可以发现地震的秘密；

反映环境问题

污染的环境会造成年轮的变化

指出地方病真相

研究树木缺少的元素从而发现当地人群缺少的元素

锯一锯

钻一钻

看一看

提纲

一、开放科学与数据科学

二、科研资源整合与关联

三、服务案例分析

四、天文信息学

定义：Wikipedia

- **天文信息学（Astro-informatics）**是研究涉及的组合的跨学科领域天文，数据科学，机器学习，信息学和信息 /通信技术。
- 主要致力于开发面向数据天文学的研究和教育的计算科学、数据科学、机器学习和统计的工具、方法和应用。
- 早期工作包括天文虚拟天文台计划中的数据发现、元数据标准开发、数据建模、天文数据字典开发、数据访问、信息检索、数据集成和数据挖掘。

https://en.wikipedia.org/wiki/Astroinformatics

大数据时代的天文学

Sky Survey Pr	Feature selection/extraction	Applied approaches	Applications in astronomy
	Feature Selection	Best First	Reducing dimension

Organization	Under community or project	Foundation Time	Chair
International Astrostatistics Association (IAA)	The International Statistical Institute (ISI)	August 2012	Joseph Hilbe
IAU Working Group in Astrostatistics and Astroinformatics	The International Astronomical Union (IAU)	August 2012	Eric Feigelson
AAS Working Group in Astroinformatics and Astrostatistics	The American Astronomical Society (AAS)	June 2012	Zeljko Ivezic
ASA Interest Group in Astrostatistics	The American Statistical Association (ASA)	March 2014	Jessi Cisnewski
LSST Informatics and Statistics Science Collaboration	The Large Synoptic Survey Telescope (LSST)	Under construction	Kirk Borne
IAA Working Group on Cosmostatistics (renamed Cosmostatistics Initiative, short for COIN)	The International Astrostatistics Association (IAA)	April 2014	Rafael de Souza

Table 4: Astrostatistics and astroinformatics organizations.

Table 2: A	Kernel Partial Least Squares (KPLS)

Table 3: Feature selection/extraction methods.

Yanxia Zhang, Yongheng Zhao. 2015. Astronomy in the Big Data Era. Data Science Journal, 14: 11, pp. 1–9, DOI: http://dx.doi.org/10.5334/dsj2015-011

天文信息学中机器学习和数据分析中相关研究问题和具体问题（2018）

- 真实大数据中的有效处理和分析
- 天文学数据中的数据挖掘和知识发现
- 天文图像数据的处理和分析
- 海量数据流的过滤技术
- 观测数据中的异常点和新颖点检测
- 天体分类和聚类分析
- 透明交互模型和算法开发
- 天文物理学模型方针和相关推理问题
- 多源异构数据集分析
- 迁移学习和特权信息学习方法
- 在循环方法中开发人工（专家）知识
- 在机器学习中系统地整合原有信息和领域知识

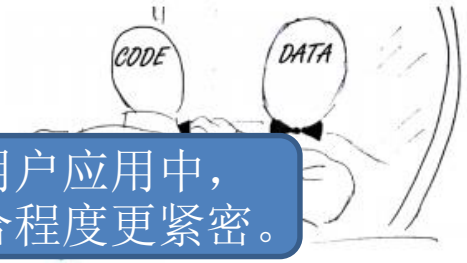
Michael Bieh, Kerstin Bunte, Guiseppe Longo, and Peter Tino. Machine Learning and Data Analysis in Astroinformatics.ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2018-2.pdf

数据驱动天文研究中的天文信息学（**挑战**）

The changing landscape of astronomical research

- **Past:** 100's to 1000's of independent distributed heterogeneous data/metadata repositories.
- **Today:** astronomical data are now accessible uniformly from federated distributed heterogeneous sources = **Virtual Observatory**.
- **Future:** astronomy is and will become even more data-intensive in the coming decade with the growth of massive data-producing sky surveys.

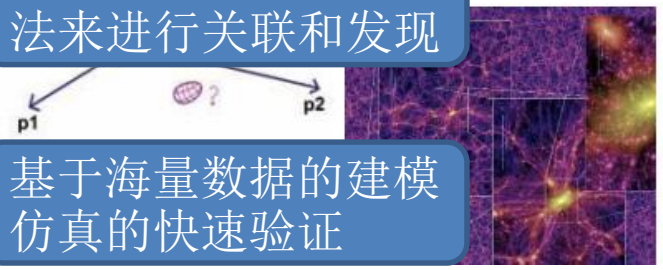
Challenge #1: it will be prohibitively difficult to transport the data to the user application. Therefore … ***SHIP THE CODE TO THE DATA !*** ***We need Distributed Data Mining methodology…***



数据很难推送到用户应用中，
代码和数据的结合程度更紧密。

Challenge #2: surveys are useful to measure and collect data from all objects present in large regions of sky, in a systematic, controlled, repeatable fashion. But … **AUTOMATIC SELF-ADAPTIVE METHODS ARE REQUIRED TO EXPLORE AND CROSS-CORRELATE THEIR DATA!**

缺乏自动自适应的方法来进行关联和发现



Challenge #3: we must be ready when huge of data will come. Mock data must be provided to ensure that data analytics methods will be compliant, efficient and scalable. Therefore … ***IMPROVE SIMULATIONS AND INFRASTRUCTURES TO MAKE INTENSIVE TESTS ON YOUR CODE!***

基于海量数据的建模仿真的快速验证

Astroinformatics in the data-driven Astronomy，2017，
https://indico.ict.inaf.it/event/535/contributions/1190/attachments/851/1204/ICT2017_Astroinformatics_Brescia.pdf

数据驱动天文研究中的天文信息学（挑战）

General Challenges in Astronomy over next decade addressable by Astroinformatics

Scalability of statistical, computational & data mining algorithms to peta- and exa- scales

Algorithms to optimize of simultaneous multi-point fitting across massive **multi-dimensional data cubes**

Petascale analytics for **visual data analysis** of massive databases (including feature detection, pattern discovery, clustering, class discovery, dimension reduction)

Rapid query, **cross-matching** and search algorithms for highly-dimensional petabyte databases

PB级数据的统计、计算、挖掘算法的扩展性

多维数据建模仿真过程中的算法优化

PB级数据的可视化分析和展示

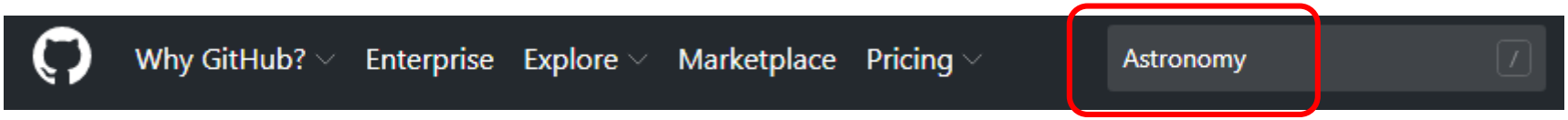
快！快！快！PB级数据中的快速查询、交叉匹配和检索算法

Astroinformatics in the data-driven Astronomy，2017，
https://indico.ict.inaf.it/event/535/contributions/1190/attachments/851/1204/ICT2017_Astroinformatics_Brescia.pdf

天文信息学年会（AstroInformatics）

- 2020（哈佛大学）：机器学习和其他高级数据分析方法，大数据集的处理，处理和分发，数据可视化，数据科学中的方法转移，天文信息学在大型调查和项目中的作用，新兴技术等。
- 2019（加州大学）：数据科学和X-信息学，天文信息学方法和应用，天文信息学大型项目，技术转移、量子计算和展望。
- 2018（德国海德堡）：科研信息化基础设施、数据挖掘与知识发现、可视化和数据探索、数据库系统和数据密集型项目和调查和时域天文学。

Github中天文学的“代码库”



Languages

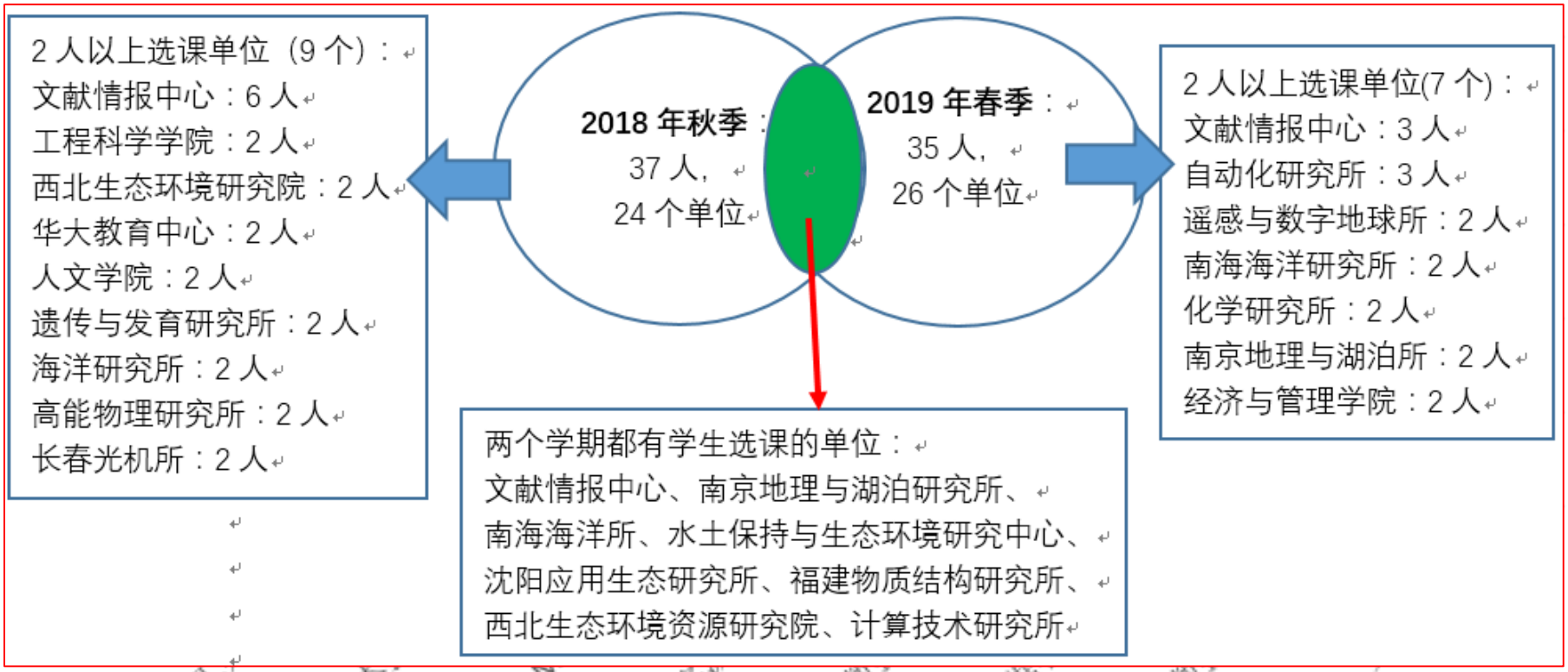
Python	1,046
JavaScript	435
Jupyter Notebook	418
HTML	224
Java	182
C++	148
C	118
TeX	112
CSS	99
C#	70

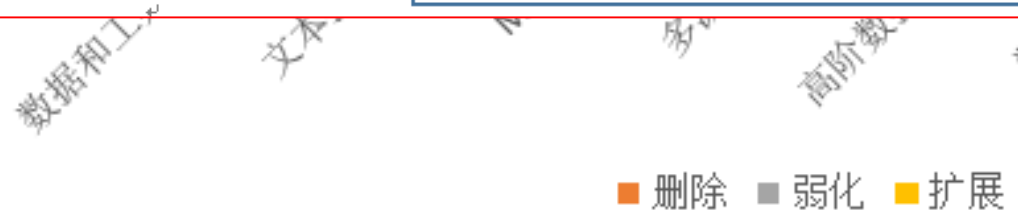
名称	简介	代码	stars
astropy/astropy	天文数据处理核心包	Python	**2.3K**
Stellarium/stellarium	利用OpenGL技术渲染实际星空图	C++	**1.5K**
jagi/meteor-astronomy	Model layer for Meteor	JavaScript	598
astroML/astroML	机器学习、统计和数据挖掘	Python	574
CelestiaProject/Celestia	实时的天空3D可视化	C++	488
sunpy/sunpy	太阳物理Python包	Python	462
skyfielders/python-skyfield	Elegant *astronomy* for Python	Python	403

关于人的问题：《科学数据管理与应用》

章节标题	学时	课程内容
第一章、基础理论	3	介绍科学数据与科研活动的紧密联系以及数据科学的特点，分析成为数据科学家的实际技能需求，通过科研信息化和开放科学框架介绍科学数据的相关概念和实践利用。
第二章、数据收集与获取	6	介绍不同来源渠道的科学数据获取和清洗方法，学习火车头、DownThemAll!下载插件、文本处理工具的使用。
第三章、数据分析与揭示	9	介绍数据库数据（MYSQL）、多媒体数据和空间数据的处理，介绍可视化技术（如百度ECharts等）、知识图谱、数据分析脚本（R和Python简介）和科研工作流工具（KNIME）的应用。
第四章、数据存储与共享	3	介绍数据整合与共享中应用的学科元数据和多媒体对象的元数据的读写操作实例，同时介绍数据共享中的FAIR原则，以及数据引用的相关情况。
第五章、数据管理规划	3	从DMP（数据管理计划）角度对实际项目或课题所需要的数据进行采集、组织、管理、存储、安全运维、长期保存和共享工作方面的考虑和设计，并介绍DMP的模板、工具和实例。
课堂实习	3	提前两周安排，使用上课讲授的知识、技能和工具，处理和交付老师给予或者学生自己提出的分组作业。
大开卷考试	3	内容涉及DMP（数据管理计划）、专业领域典型数据平台介绍、专业领域元数据标准介绍、学科史或者人文素材分析等。
合计	30	

学生对课程的反馈（35人）





选课学生认为的课程的修改建议

学生实践题目

2018年秋季

- SQL在生物实验数据管理中的应用
- 春秋战国以来我国疆域变化（CHGIS）
- 电信行业套餐预测
- 华为专利申请分析
- 污染源的化工厂在全国的分布图
- 装配式建筑企业在全国的分布图

2019年春季

- 《红楼梦》人物关系图谱分析
- 电影和电影明星关系分析
- 房价与就业率的关系分析
- 骑行信息与热门健身运动分析
- 全国污染物与经济增长分析

选课学生对课程的具体建议

感兴趣的知识点	学习中的困难和没掌握的知识	需深入学习的知识点	希望增加的知识点	可弱化知识点
1) 数据挖掘	1) 数据存储与共享部分讲的有点少，希望扩展	1) MYSQL 数据库和 Python	1) 数据在科研中的位置或使用逻辑	1) MySQL 语句
2) 可视化技术	2) 没能及时在课下练习老师课上讲授的内容	2) 高阶数据挖掘和分析	2) 希望今后可以增加实习时间，一边能及时掌握	2) 遥感数据处理
3) 数据的获取方式，如火车头	3) 数据平台使用联系少	3) 数据管理规划	3) 分布或数据分析	3) 文本抓取和清洗
4) 数据分析，数据据存储与关系	4) 文本抓取和清洗	4) SQL 语句、高阶数据挖掘	4) 数据类型介绍、数据分析过程	4) 一次性抓取数据工具完全可以用爬虫替代
5) 数据管理	5) 多源数据中的 GIS 部分	5) 可视化技术	5) 爬虫	5) MYSQL 和 SQL 操作
6) 工具平台推荐	6) 高阶数据挖掘	6) 数据库和数据分析	6) 编程语言	6) 基础理论
7) GIS	7) 火车头	7) R 语言与 Python 处理	7) SQL 进一步强化	7) 数据存储与共享
8) 多源数据操作以及高阶数据挖掘	8) 多源数据操作	8) Python 脚本、火车头采集器等以后科研、生活会用到的工具	8) 基础理论	8) 多媒体、GIS
9) 数据查询	9) 代码不了解	9) SQL、各数据平台	9) 高阶网络爬虫	9) 数据和工具平台介绍可以适当减弱
10) 多媒体/GIS/可视化技术；	10) 可视化工具的学习	10) 高阶数据挖掘，多媒体，GIS，可视化技术	10) 数据分析	10) 数据存储/协议
11) 数据库的建设、维护和管理、数据分析	11) 数据管理规划，因为还没有实践经历	11) MySQL	11) Python 语言 R 语言在数据挖掘中的应用	11) 高阶数据挖掘
12) SQL 和 SQL 操作	12) 数据存储与共享	12) R 分析	12) xml/php 等前端知识	12) 编程语言
13) 高阶数据挖掘	13) 软件太多，讲得太快，许多未掌握	13) 文本抓取和清洗以及数据存储与共享	13) 对数据类型与分析，视频剪辑处理，数据平台与资源获取	13) 编程
14) 数据抓取，数据库实践	14) 各个软件操作与用途讲的太快了，跟不上老师进度	14) 数据工具平台使	14) R/PYTHON 可视化	14) 数据和工具平台介绍以及数据管理计划
15) 可以学习大数据分析技术，并且可以学到非技术手段的工具	15) 火车头以及 MySQL/SQL 操作语句		15) 增加一些文本编程方式类基础的内容	15) 以为不必学火车头，我觉得爬取效果不理想，学起来更复杂，不如用 python 代码爬简单
16) SQL\Python	16) 有关 python 方面			16) DMP/数据分析
	17) 编程语言			
	18) 文本抓取水平还达不到应用到科研实践的需求			

小结和讨论

1. 数据是有生命力的，数据密集型科研活动会催生新的发现，**X-信息学**的前景越来越好，要求也会越来越高。
2. 挑战与机遇并存，从开放科学的“**生态系统**”角度去理解和挖掘服务。未来是一个“组装时代”，组装元素包括：学科研究、数据、人、财、物、技术。
3. 人（**数据科学家**）的能力和作用至关重要，一定要实战中才能提升自己。勇于淘汰旧我，走出舒适区，增值自己。
4. **多样化的服务方式**来适配个性化的服务场景。
5. “**资源+知识+技术**”的新型服务和介入模式。

中国科学院文献情报中心
许哲平：xuzp@mail.las.ac.cn