COMMITTEE ON DATA
CODATA
INTERNATIONAL
SCIENCE COUNCIL

International
Science Council

Data interoperability in China: practice and challenges

Lianglin Hu (HULL@cnic.cn)

CNIC.CAS/NBSDC.CN/GOSC DataIO WG

2022.10.10

Abbreviated terms

Abbreviated term	All words term
CAAS	Chinese Academy of Agricultural Sciences
CAS	Chinese Academy of Sciences
CAS Project	Scientific Data Project in CAS, started in 1986
CNIC	Computer Network Information Center, CAS
CSTCloud	China Science and Technology Cloud, building and running by CNIC
DataIO	Data interoperability
EIF	European Interoperability Framework
GOSC	Global Open Science Cloud
MOST	Ministry of Science and Technology, China
MOST Project	National Scientific Data Sharing Project, started in 2001 by MOST
National Measure	Measures for Scientific Data Management, China
NBSDC	National Basic Science Data Center, China

Contents

- Two projects
- DataIO practice
- Challenges

Scientific Data Project in CAS

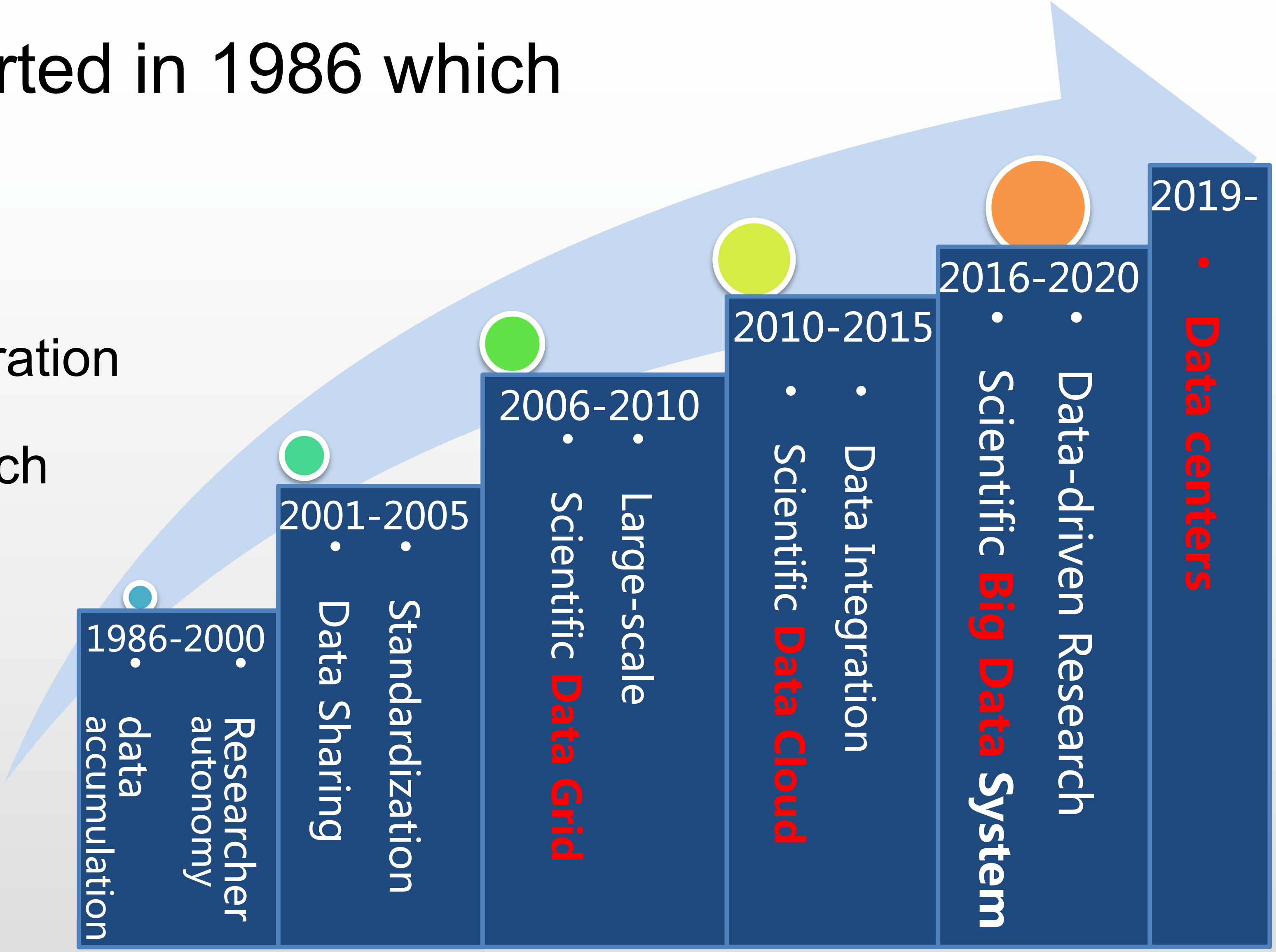
- A Long-term mission started in 1986 which funded by CAS
  - 60+ institutes involved
  - long-term, large-scale collaboration
  - data from research, for research

科技進步獎
證書

为表彰在促进科学技术进步工作中做出重大贡献者，特颁发国家科技进步奖证书，以资鼓励。

[illegible]

宋健



1986-2000
- Researcher autonomy
- data accumulation

2001-2005
- Standardization
- Data Sharing

2006-2010
- Large-scale
- Scientific **Data Grid**

2010-2015
- Data Integration
- Scientific **Data Cloud**

2016-2020
- Data-driven Research
- Scientific **Big Data System**

2019-
- **Data centers**

National Scientific Data Sharing Project

China has long attached importance to the management and sharing of scientific data

- In 2001, "implement the scientific data sharing project and strengthen the national science and technology innovation capability" was put forward. In December, the Ministry of Science and Technology launched the first scientific data sharing pilot for meteorological science data sharing.
- In December 2002, the Ministry of Science and Technology launched 8 other shared pilots: agriculture, forestry, hydrology and water resources, earthquakes, surveying and mapping, earth system science, sustainable development, and rural modern science and technology information.
- In 2003, with the support of the Ministry of Finance, the Ministry of Science and Technology set up the National Science and Technology Infrastructure (NSTI). The scientific data sharing project was included as an important component in the construction of the NSTI.
- In 2009, the first batch of scientific data sharing projects were accepted and transferred to the operational service phase. So far, eight scientific data sharing platforms have been identified by the NSTI.
- In 2017, the "13th Five-Year National Science and Technology Innovation Base and Special Conditions for Capacity Building" integrated the Science Data Centers into the National Science and Technology Innovation Base of Basic and Conditional Support.

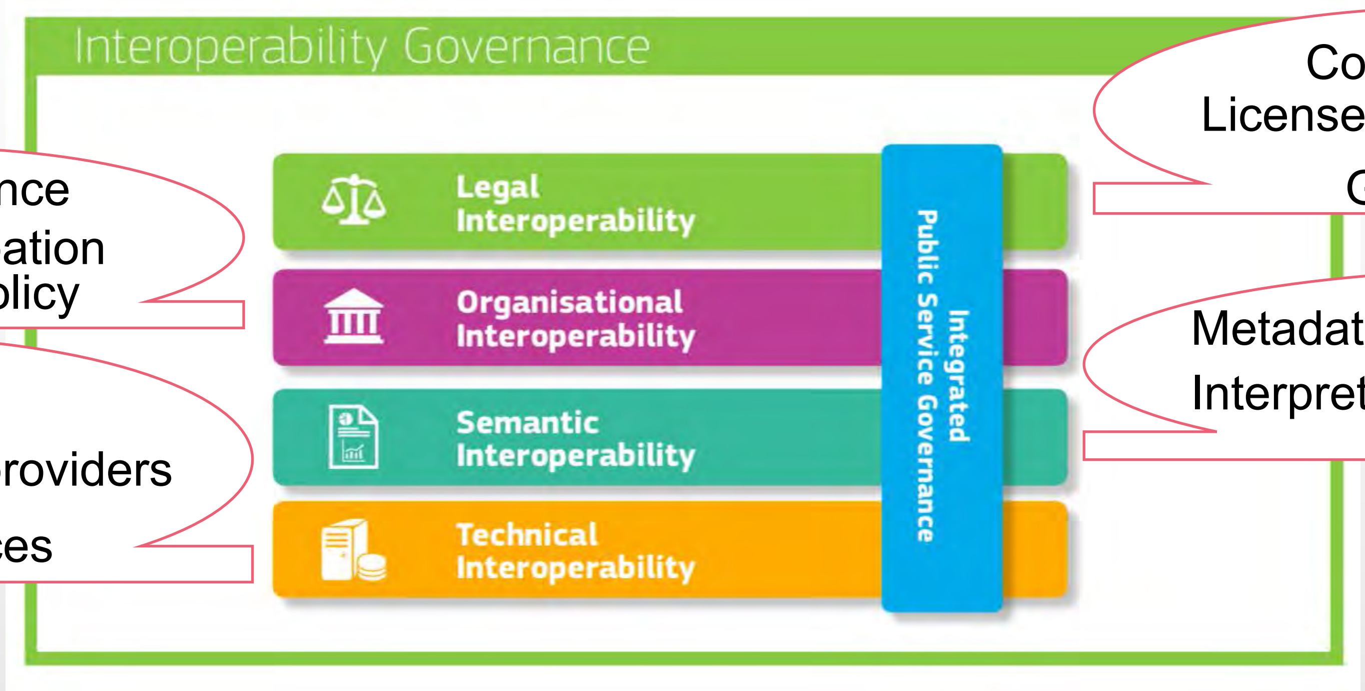
Summary

1. **Both are long-run project**
    - CAS project - 36 years
    - MOST project - 21 years
2. **Both are national level indeed**
    - CAS project was approved by the State Planning Commission in 1986
3. **Both support the best scientific data working teams in China**
    - Some institutions of CAS project are also participating body of MOST project
    - MOST project includes a number of institutions out of CAS
4. **Both lay a good foundation**
    - data, standards, softwares, platforms, services and so on
    - policy basis for National Measures
5. **Both focus on data centers now**

Contents

- ➢Two projects
- ➢DataIO practice
- ➢Challenges

European Interoperability Framework


Interoperability Governance
Legal Interoperability
Organisational Interoperability
Semantic Interoperability
Technical Interoperability
Integrated Public Service Governance

Governance
Rules of participation
Policy

Frameworks
Infrastructure providers
Services

Copyright
Licenses
Data protection
GDPR

Metadata & Ontologies
Interpretation of meaning & structure

Source: The European Interoperability Framework four levels of interoperability

From Milan's presenation: **Data interoperability in Europe: where we are and where we are going?**

L1： National policy - Measures for Scientific Data Management

- drafted by the Ministry of Science and Technology
- issued by the General Office of the State Council on March 17, 2018
- establish a guideline for China's scientific data management and sharing

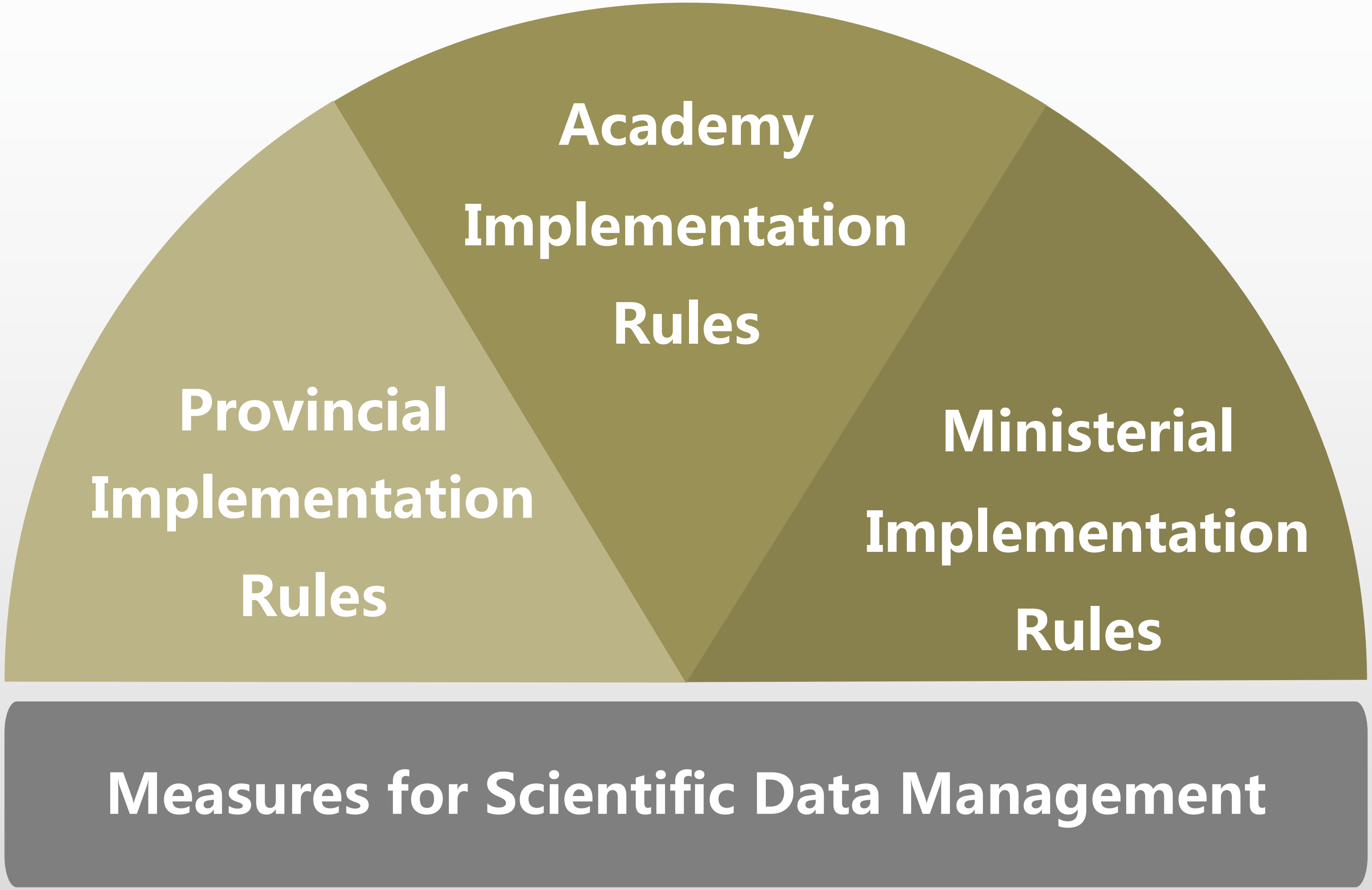
1 Deeply follow the development trend of scientific data in the era of big data. Fully draw on advanced experience and mature practices at home and abroad to strengthen the life cycle management of scientific data.

2 Prioritize ensuring data security. Scientific data sharing should be based on security and controllability, and strengthen the data resources protection capacity.

3 Vigorously promote the open sharing of scientific data resources. Adhere to the principle of "openness being the norm and non-openness being the exception", especially the sharing and opening of the scientific data supported by the state fund.

4 Focus on the weak links in the work of scientific data in China. Propose measures from the main responsibility, intellectual property rights, and the aggregation mechanism.

Policy interoperability



Academy Implementation Rules

Provincial Implementation Rules

Ministerial Implementation Rules

Measures for Scientific Data Management

Policy interoperability

- **Provincial Implementation Rules**

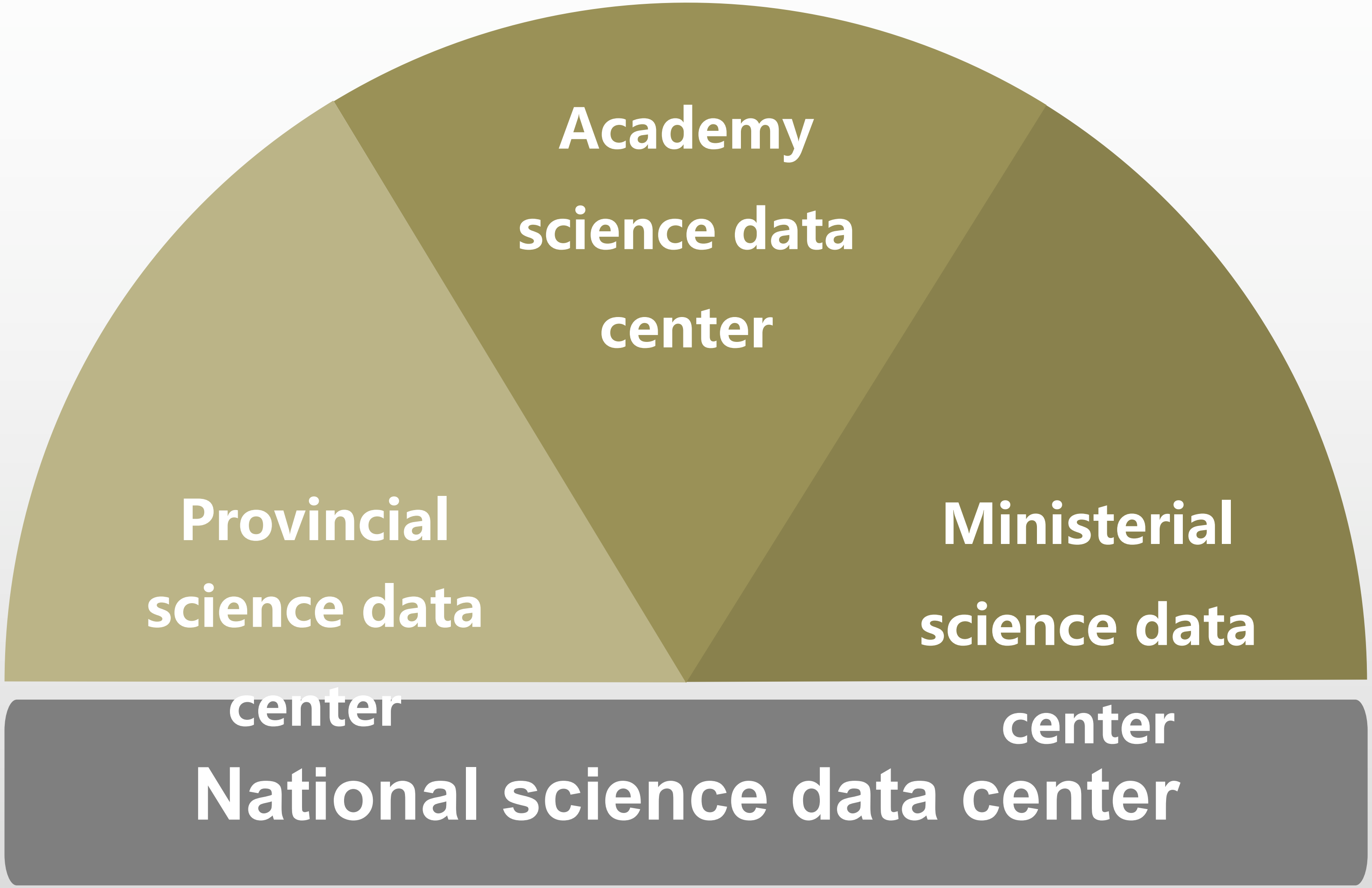
  (14+1)

  - Shaanxi province
  - Heilongjiang province
  - Gansu province
  - Yunnan province
    - kunming city
  - Hubei province
  - Anhui province
  - Inner Mongolia Autonomous Region
  - Jilin province
  - Guangxi province
  - Chongqing city
  - Jiangsu province
  - Hainan province
  - Shandong province
  - Sichuan province

Policy interoperability

- **Academy Implementation Rules**
  - Chinese Academy of Sciences
  - Chinese Academy of Agricultural Sciences
- **Ministerial Implementation Rules**
  - Ministry of Transport

L2：Organisational interoperability


Academy science data center
Provincial science data center
Ministerial science data center
National science data center

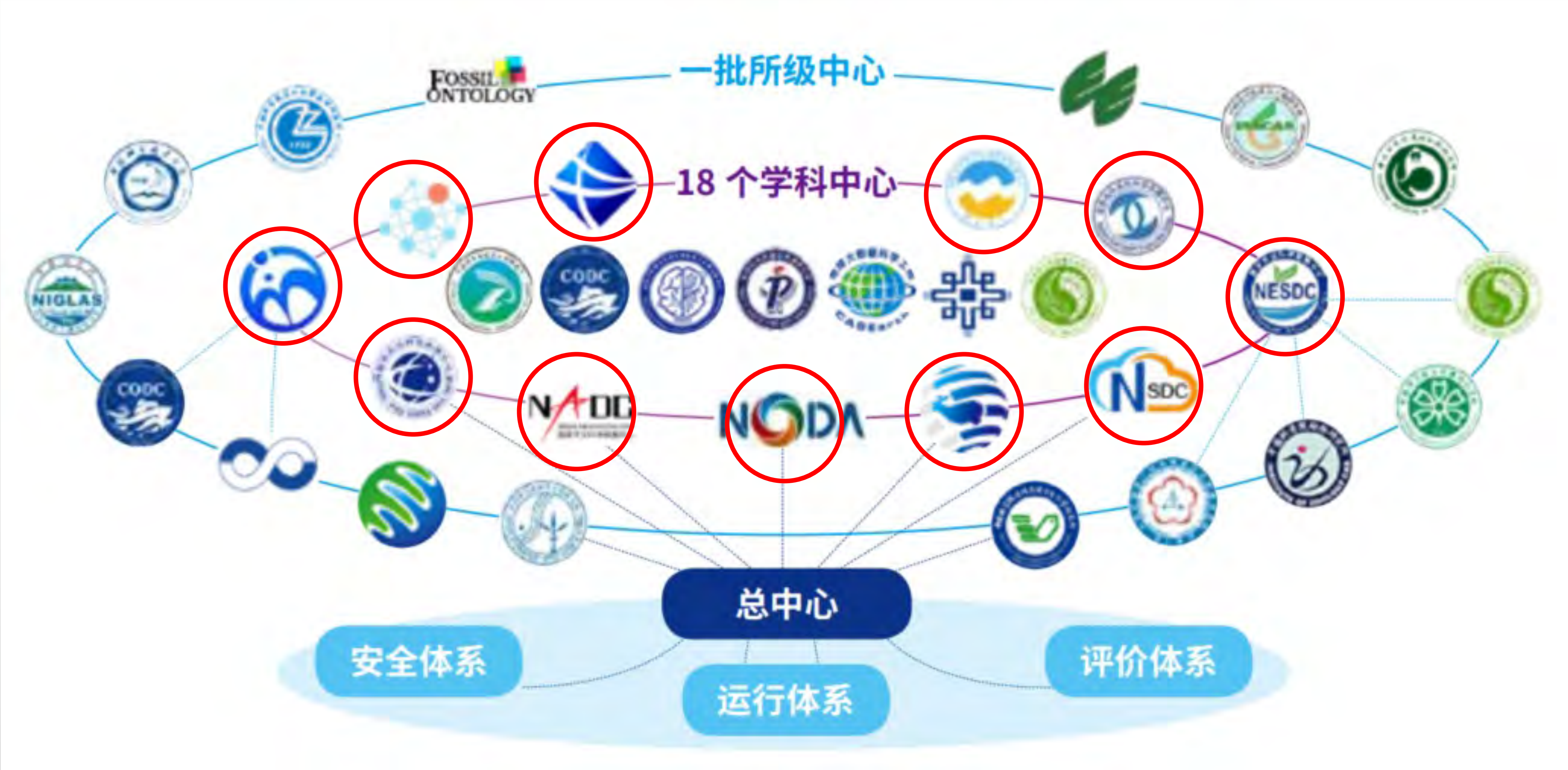
L2：
Organisational interoperability

●National science data center

No.	National Scientific Data Center
1	National High Energy Physics Science Data Center
2	National Genomics Data Center
3	National Microbiology Data Center
4	National Space Science Data Center
5	National Astronomical Data Center
6	National Earth Observation Data Center
7	National Polar Science Data Center
8	National Qinghai Tibetan Plateau Data Center
9	National Ecosystem Science Data Center
10	National Corrosion & Protection Data Center
11	National Cryosphere Desert Data Center
12	National Metrology Data Center
13	National Earth System Science Data Center
14	National Population Health Data Center
15	National Basic Science Data Center
16	National Agricultural Science Data Center
17	National Forestry and Grassland Data Center
18	National Meteorological Data Service Center
19	National Earthquake Science Data Center
20	National Marine Data Center

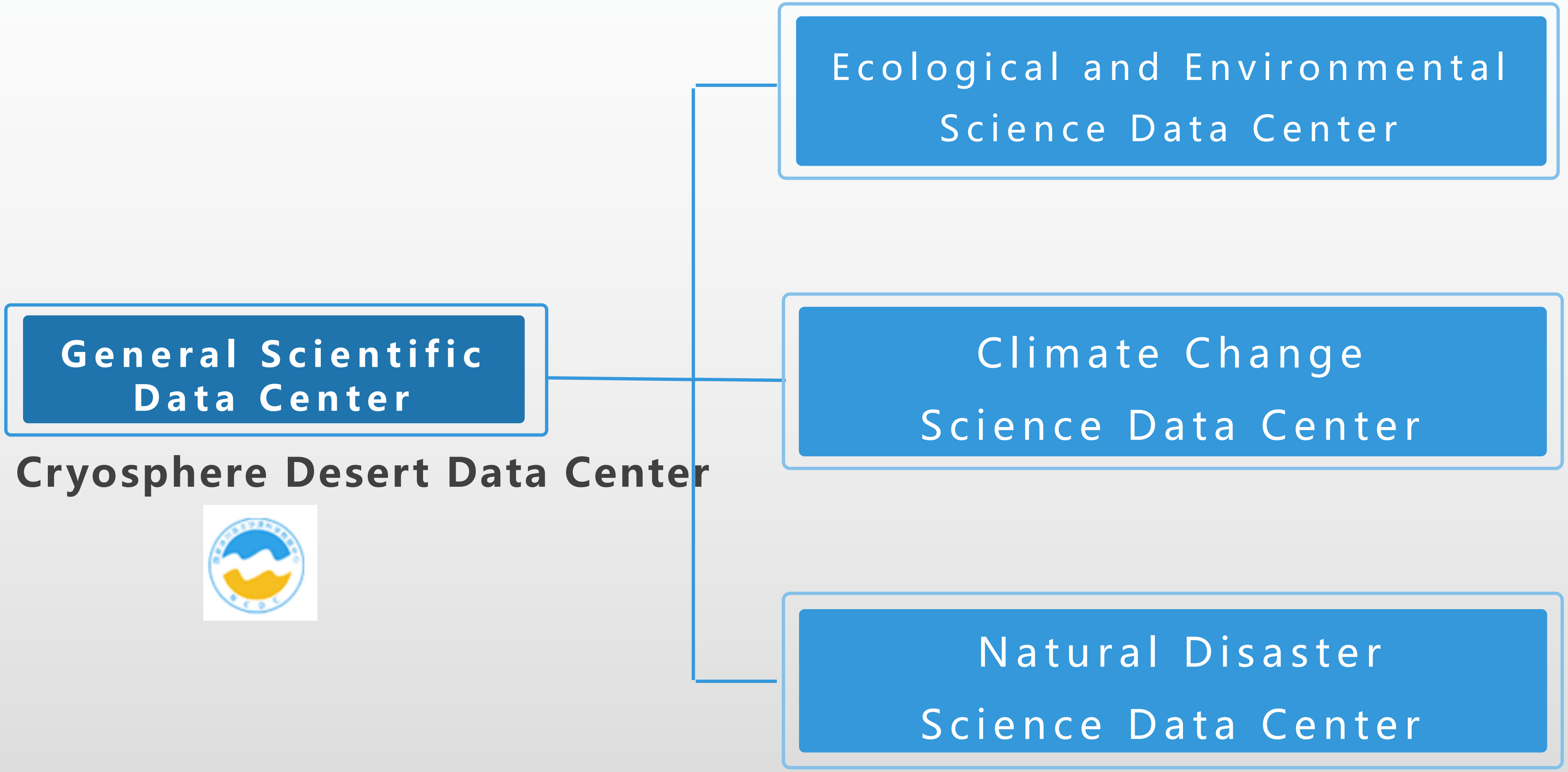
Organisational interoperability

- **CAS scientific data center**



Organisational interoperability

- **Gansu provincial science data center**



General Scientific Data Center

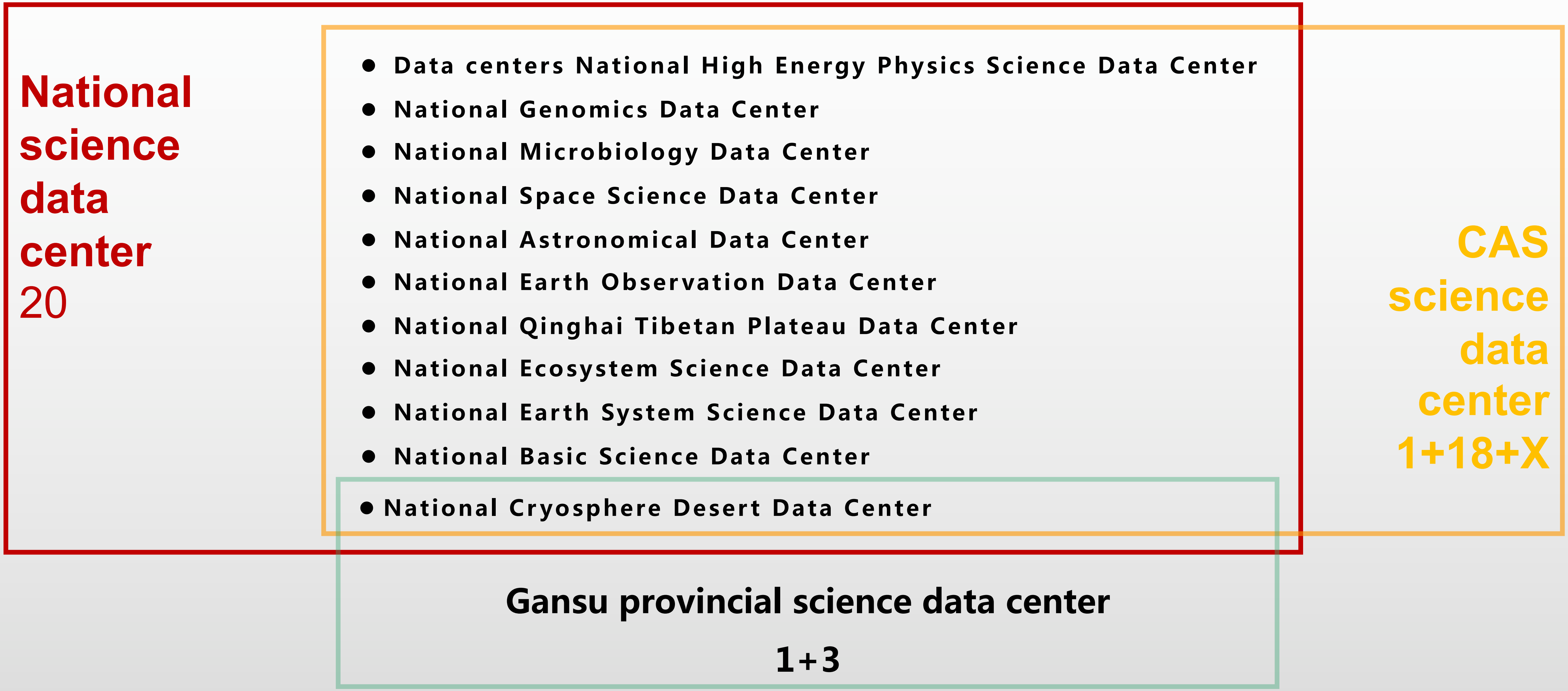
National Cryosphere Desert Data Center

Ecological and Environmental Science Data Center

Climate Change Science Data Center

Natural Disaster Science Data Center

Organisational interoperability



National science data center 20

- Data centers National High Energy Physics Science Data Center
- National Genomics Data Center
- National Microbiology Data Center
- National Space Science Data Center
- National Astronomical Data Center
- National Earth Observation Data Center
- National Qinghai Tibetan Plateau Data Center
- National Ecosystem Science Data Center
- National Earth System Science Data Center
- National Basic Science Data Center
- National Cryosphere Desert Data Center

CAS science data center 1+18+X

Gansu provincial science data center

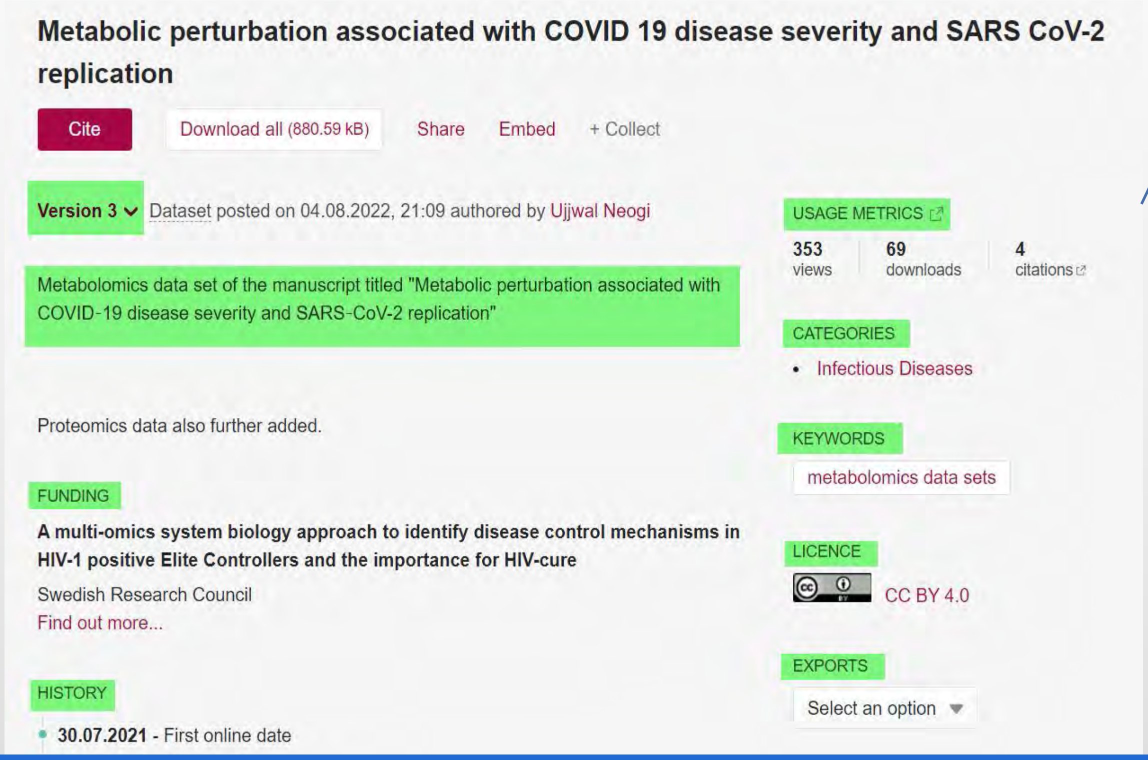
1+3

L3 : Semantic interoperability

- Metadata
  - Data about data

dataset metadata

dataset includes data and its' metadata



Metabolic perturbation associated with COVID 19 disease severity and SARS CoV-2 replication

Cite | Download all (880.59 kB) | Share | Embed | + Collect

Version 3 Dataset posted on 04.08.2022, 21:09 authored by Ujjwal Neogi

Metabolomics data set of the manuscript titled "Metabolic perturbation associated with COVID-19 disease severity and SARS-CoV-2 replication"

Proteomics data also further added.

FUNDING

A multi-omics system biology approach to identify disease control mechanisms in HIV-1 positive Elite Controllers and the importance for HIV-cure

Swedish Research Council

Find out more...

HISTORY

- 30.07.2021 - First online date

USAGE METRICS

353 views | 69 downloads | 4 citations

CATEGORIES

- Infectious Diseases

KEYWORDS

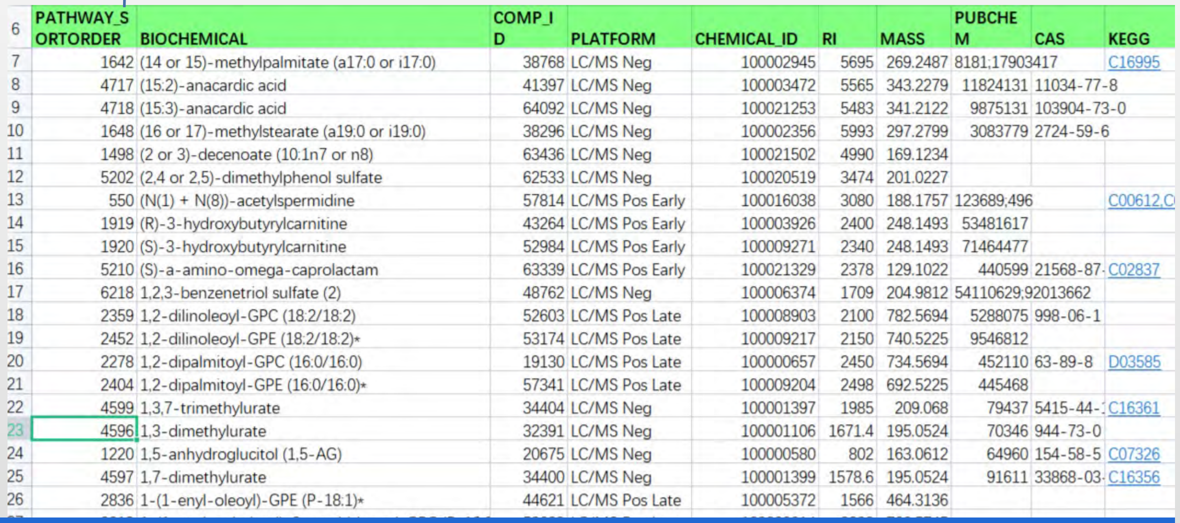
metabolomics data sets

LICENCE

CC BY 4.0

EXPORTS

Select an option

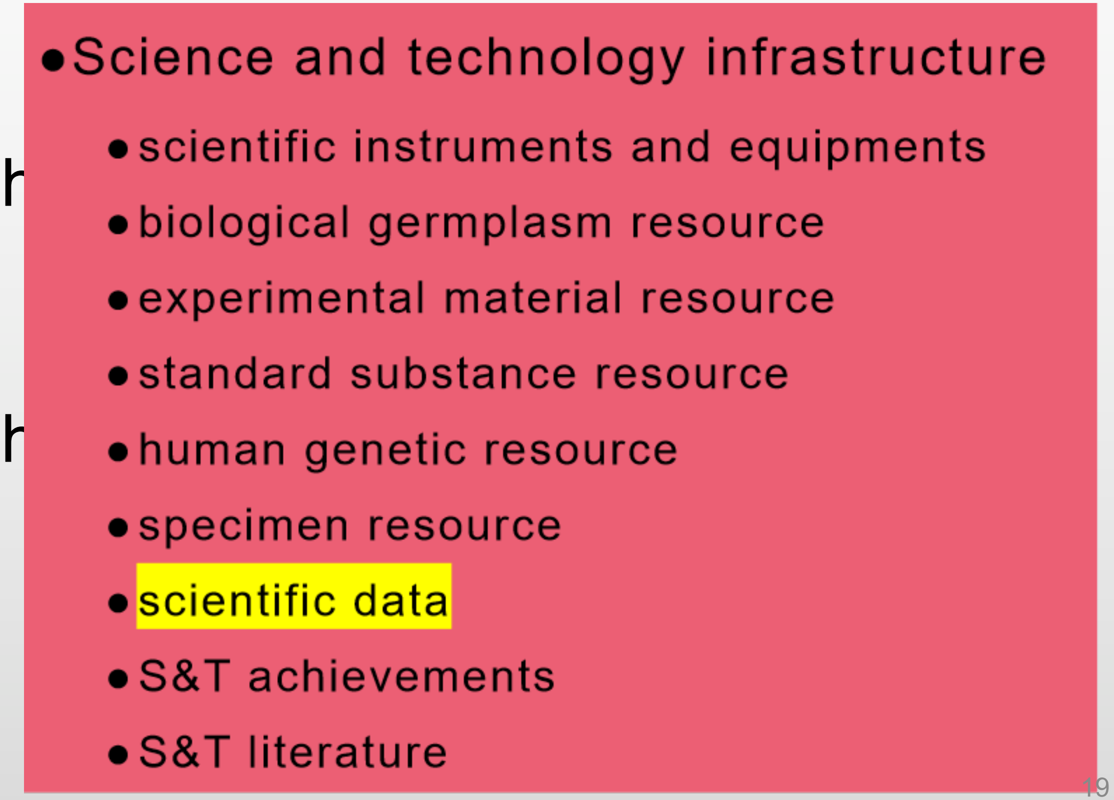


6	PATHWAY_SORTORDER	BIOCHEMICAL	COMP_ID	PLATFORM	CHEMICAL_ID	RI	MASS	PUBCHEM	CAS	KEGG
7	1642	(14 or 15)-methylpalmitate (a17:0 or i17:0)	38768	LC/MS Neg	100002945	5695	269.2487	8181;17903417		C16995
8	4717	(15:2)-anacardic acid	41397	LC/MS Neg	100003472	5565	343.2279	11824131	11034-77-8	
9	4718	(15:3)-anacardic acid	64092	LC/MS Neg	100021253	5483	341.2122	9875131	103904-73-0	
10	1648	(16 or 17)-methylstearate (a19:0 or i19:0)	38296	LC/MS Neg	100002356	5993	297.2799	3083779	2724-59-6	
11	1498	(2 or 3)-decenoate (10:1n7 or n8)	63436	LC/MS Neg	100021502	4990	169.1234			
12	5202	(2,4 or 2,5)-dimethylphenol sulfate	62533	LC/MS Neg	100020519	3474	201.0227			
13	550	(N(1) + N(8))-acetylspermidine	57814	LC/MS Pos Early	100016038	3080	188.1757	123689;496		C00612,C
14	1919	(R)-3-hydroxybutyrylcarnitine	43264	LC/MS Pos Early	100003926	2400	248.1493	53481617		
15	1920	(S)-3-hydroxybutyrylcarnitine	52984	LC/MS Pos Early	100009271	2340	248.1493	71464477		
16	5210	(S)-a-amino-omega-caprolactam	63339	LC/MS Pos Early	100021329	2378	129.1022	440599	21568-87-	C02837
17	6218	1,2,3-benzenetriol sulfate (2)	48762	LC/MS Neg	100006374	1709	204.9812	54110629;92013662		
18	2359	1,2-dilinoleoyl-GPC (18:2/18:2)	52603	LC/MS Pos Late	100008903	2100	782.5694	5288075	998-06-1	
19	2452	1,2-dilinoleoyl-GPE (18:2/18:2)*	53174	LC/MS Pos Late	100009217	2150	740.5225	9546812		
20	2278	1,2-dipalmitoyl-GPC (16:0/16:0)	19130	LC/MS Pos Late	100000657	2450	734.5694	452110	63-89-8	D03585
21	2404	1,2-dipalmitoyl-GPE (16:0/16:0)*	57341	LC/MS Pos Late	100009204	2498	692.5225	445468		
22	4599	1,3,7-trimethylurate	34404	LC/MS Neg	100001397	1985	209.068	79437	5415-44-	C16361
23	4596	1,3-dimethylurate	32391	LC/MS Neg	100001106	1671.4	195.0524	70346	944-73-0	
24	1220	1,5-anhydroglucitol (1,5-AG)	20675	LC/MS Neg	100000580	802	163.0612	64960	154-58-5	C07326
25	4597	1,7-dimethylurate	34400	LC/MS Neg	100001399	1578.6	195.0524	91611	33868-03-	C16356
26	2836	1-(1-enyl-oleoyl)-GPE (P-18:1)*	44621	LC/MS Pos Late	100005372	1566	464.3136			

https://figshare.com/articles/dataset/Metabolic_perturbation_associated_with_COVID_19_disease_severity_and_SARS_CoV-2_replication/13336862?file=28993998

Semantic interoperability -- dataset metadata

- **National Standards for dataset**
  - GB/T 30523-2014,Science and technology infrastructure
    —Resource core metadata
  - GB/T 31073-2014,Science and tech
    —Core metadata of service
  - GB/T 39913-2021,Science and tech
    —User metadata



- Science and technology infrastructure
  - scientific instruments and equipments
  - biological germplasm resource
  - experimental material resource
  - standard substance resource
  - human genetic resource
  - specimen resource
  - scientific data
  - S&T achievements
  - S&T literature

Semantic interoperability -- dataset metadata

●Resource core metadata elements

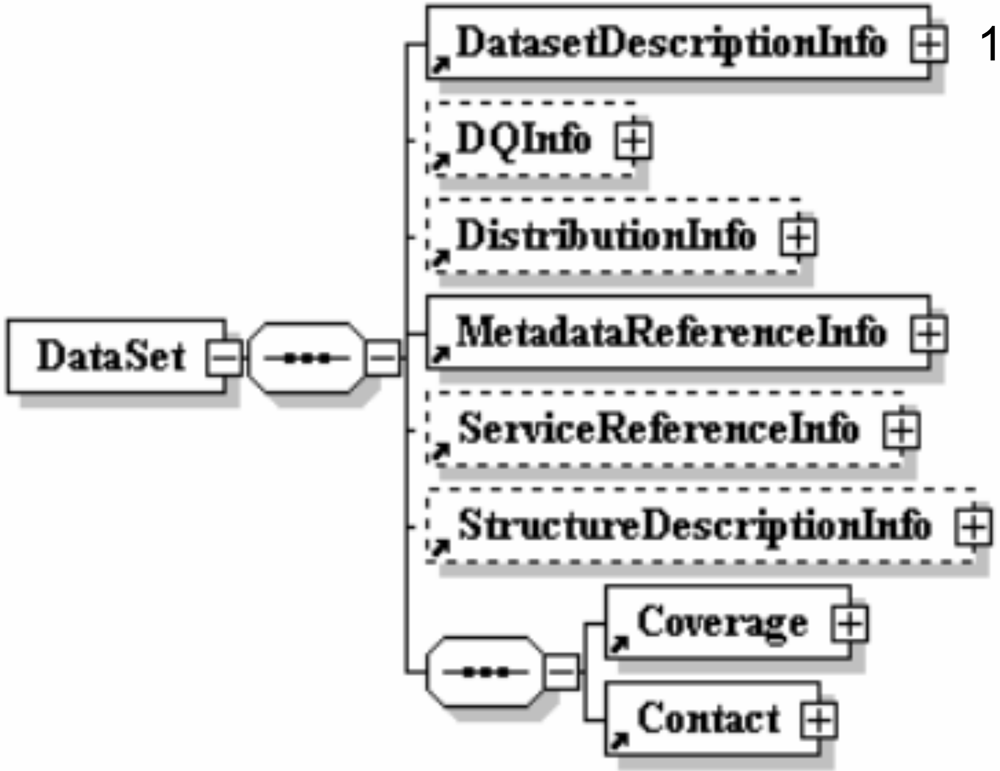
- ResourceIdentifier
- ResourceTitle
- ResourceCategory
- Keyword
- Description
- AccessConstraints
- DateOfUpdate
- PointOfContact
- OnlineAddress

□scientific instruments and equipments
□biological germplasm resource
□experimental material resource
□standard substance resource
□human genetic resource
□specimen resource
□scientific data
□S&T achievements
□S&T literature

All 20 national science data centers publish datasets based on this metadata standard.

Semantic interoperability -- dataset metadata

- **CAS Scientific Data Core metadata @ CAS project**



16 elements

2 elements

7 elements

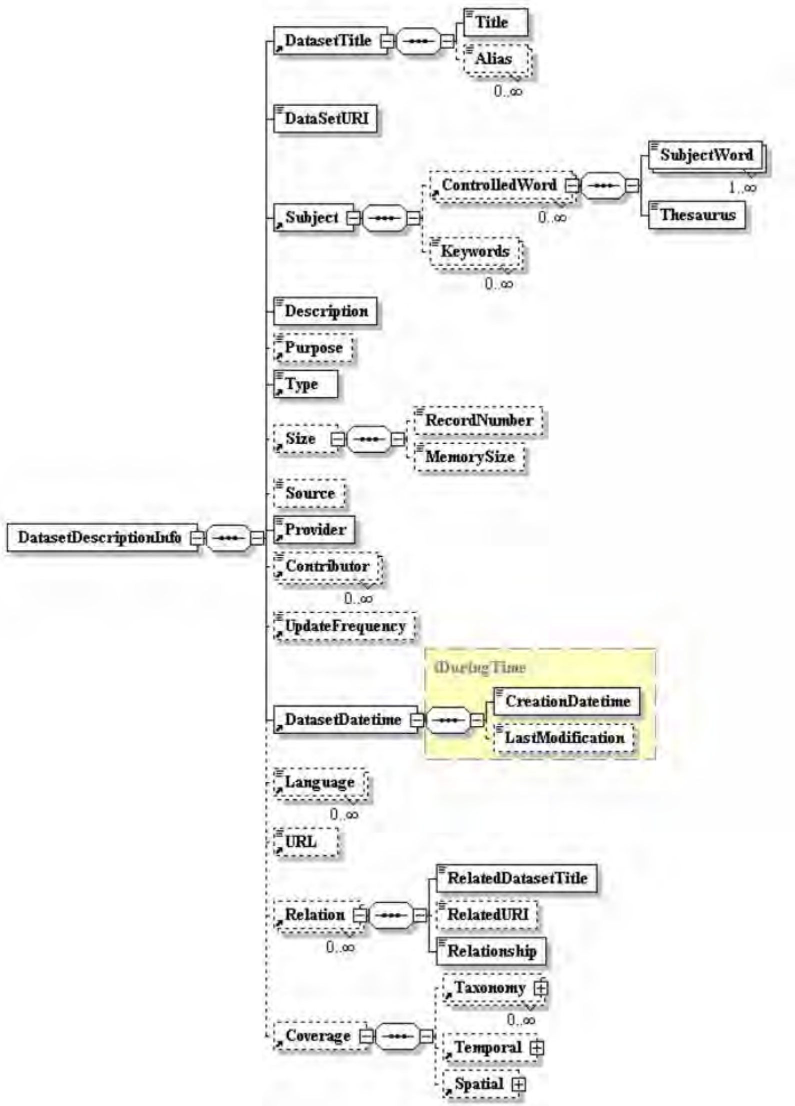
4 elements

2 elements

3 elements

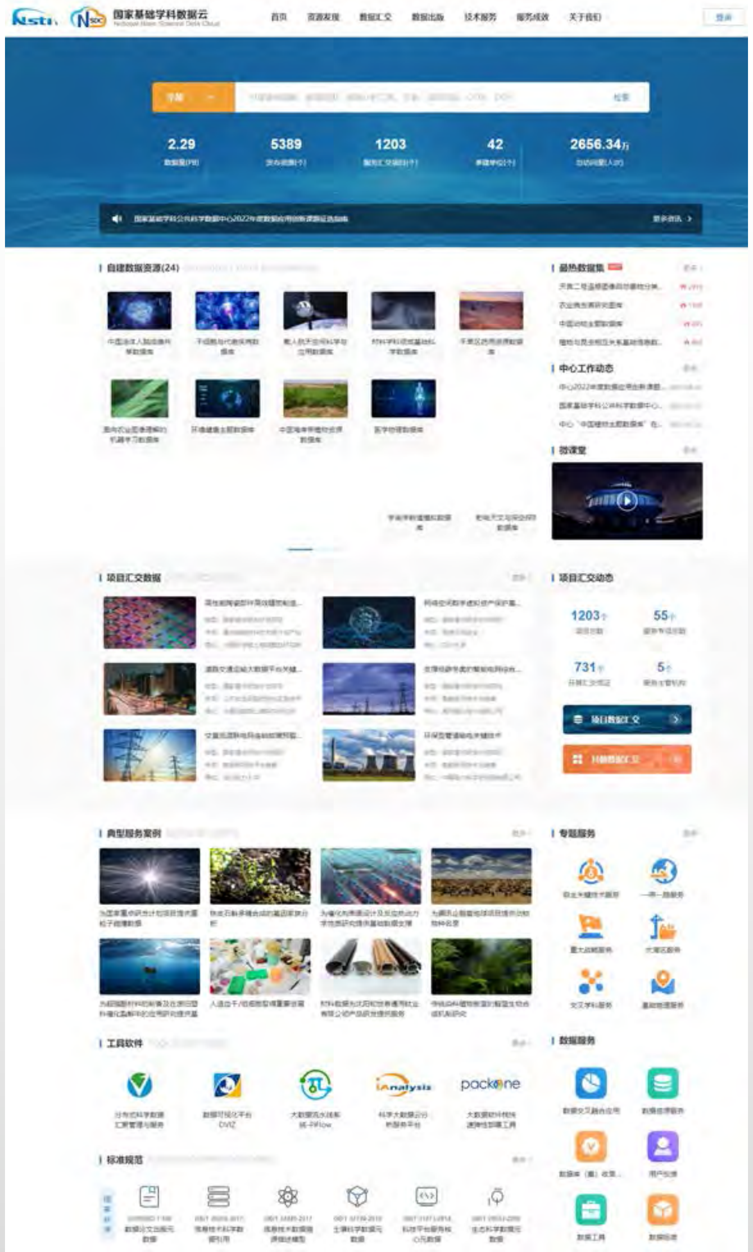
3 elements

4 elements



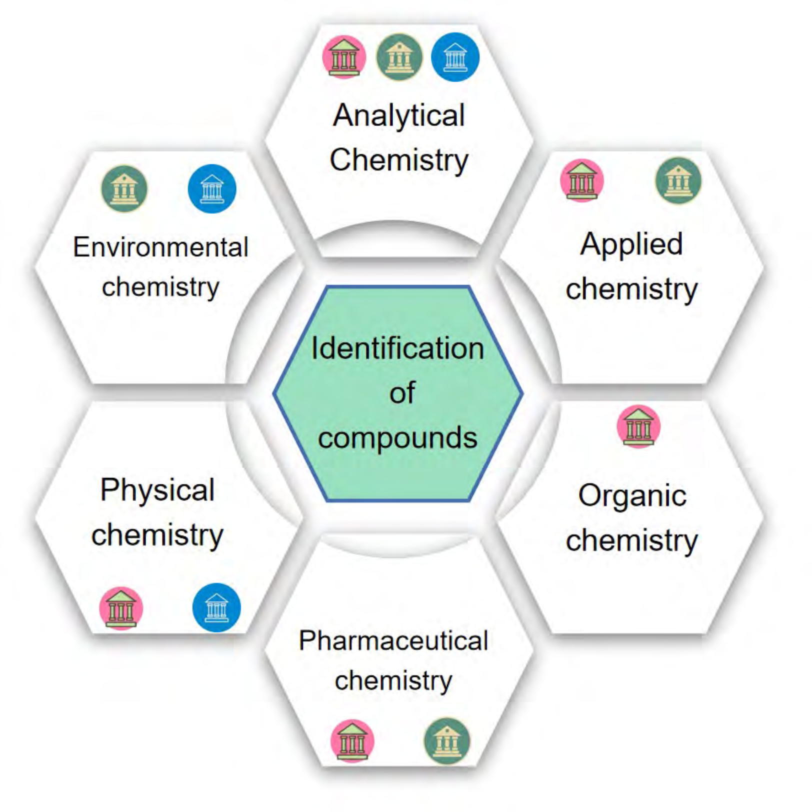
Semantic interoperability -- NBSDC Data

- **Data(set) resources**
  - Physics
  - Chemical
  - Material
  - Plant
  - Zoology
  - Information Science
  - ……



Semantic interoperability -- Data

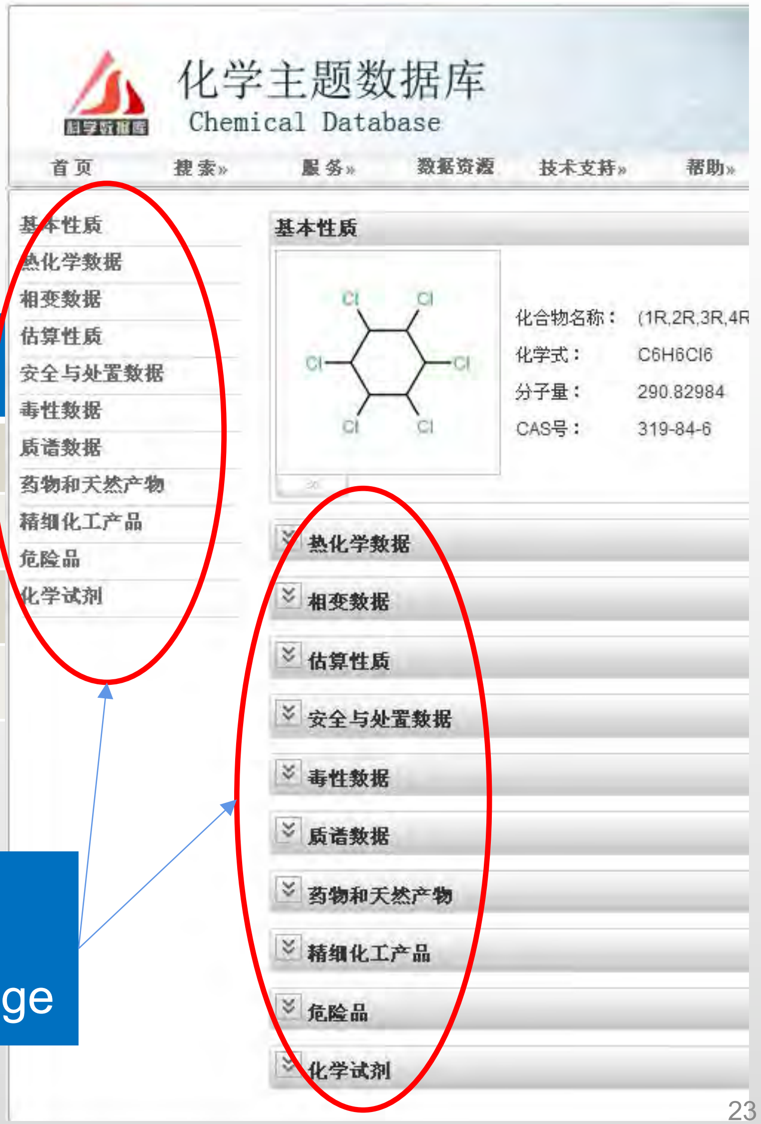
●Example： Chemical Database @NBSDC



- Identification Mapping table

IDtype	...	DataType	
CAS RN	...	varchar(10)	
InChIKey	...	varchar(27)	
SRN	...	Long(10)	
id	...	varchar(50)	

- Data from 3 institutes
- Collaborate services on the same page



化学主题数据库
Chemical Database

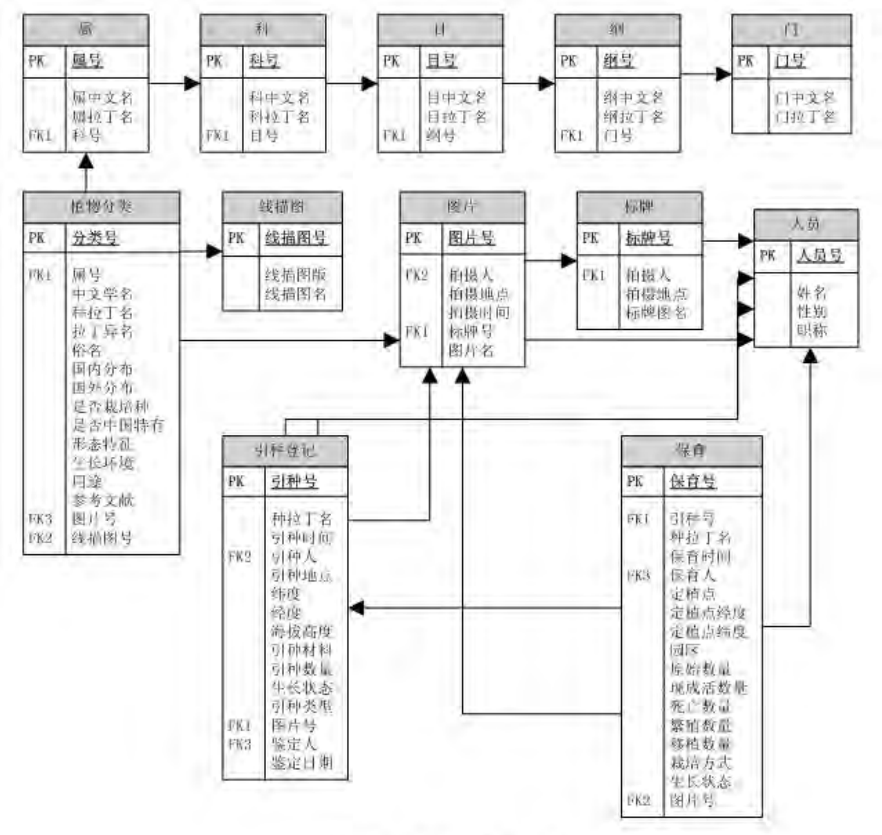
基本性质

化合物名称： (1R,2R,3R,4R
化学式： C6H6Cl6
分子量： 290.82984
CAS号： 319-84-6

热化学数据
相变数据
估算性质
安全与处置数据
毒性数据
质谱数据
药物和天然产物
精细化工产品
危险品
化学试剂

Semantic interoperability -- Data

- **Example： Plant Collection DB @NBSDC**



属	
PK	属号
	属中文名 属拉丁名
FK1	科号

科	
PK	科号
	科中文名 科拉丁名
FK1	目号

目	
PK	目号
	目中文名 目拉丁名
FK1	纲号

纲	
PK	纲号
	纲中文名 纲拉丁名
FK1	门号

门	
PK	门号
	门中文名 门拉丁名

植物分类	
PK	分类号
FK1	属号
	中文学名 种拉丁名 拉丁异名 俗名 国内分布 国外分布 是否栽培种 是否中国特有 形态特征 生长环境 用途 参考文献
FK3	图片号
FK2	线描图号

线描图	
PK	线描图号
	线描图版 线描图名

图片	
PK	图片号
FK2	拍摄人 拍摄地点 拍摄时间
FK1	标牌号 图片名

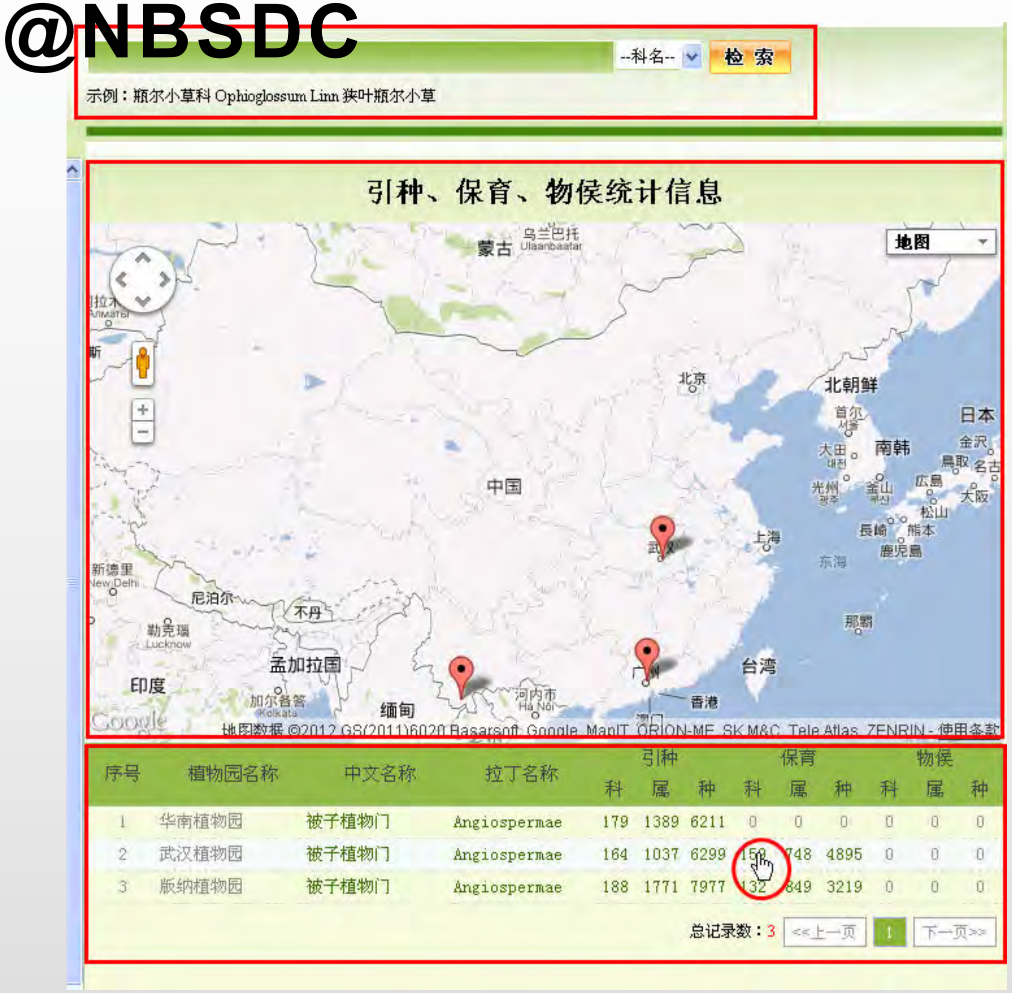
标牌	
PK	标牌号
FK1	拍摄人 拍摄地点 标牌图名

人员	
PK	人员号
	姓名 性别 职称

引种登记	
PK	引种号
	种拉丁名 引种时间
FK2	引种人 引种地点 纬度 经度 海拔高度 引种材料 引种数量 生长状态 引种类型
FK1	图片号
FK3	鉴定人 鉴定日期

保育	
PK	保育号
FK1	引种号 种拉丁名 保育时间
FK3	保育人 定植点 定植点经度 定植点纬度 园区 原始数量 现成活数量 死亡数量 繁殖数量 移植数量 栽培方式 生长状态
FK2	图片号

common data model



示例：瓶尔小草科 Ophioglossum Linn 狭叶瓶尔小草

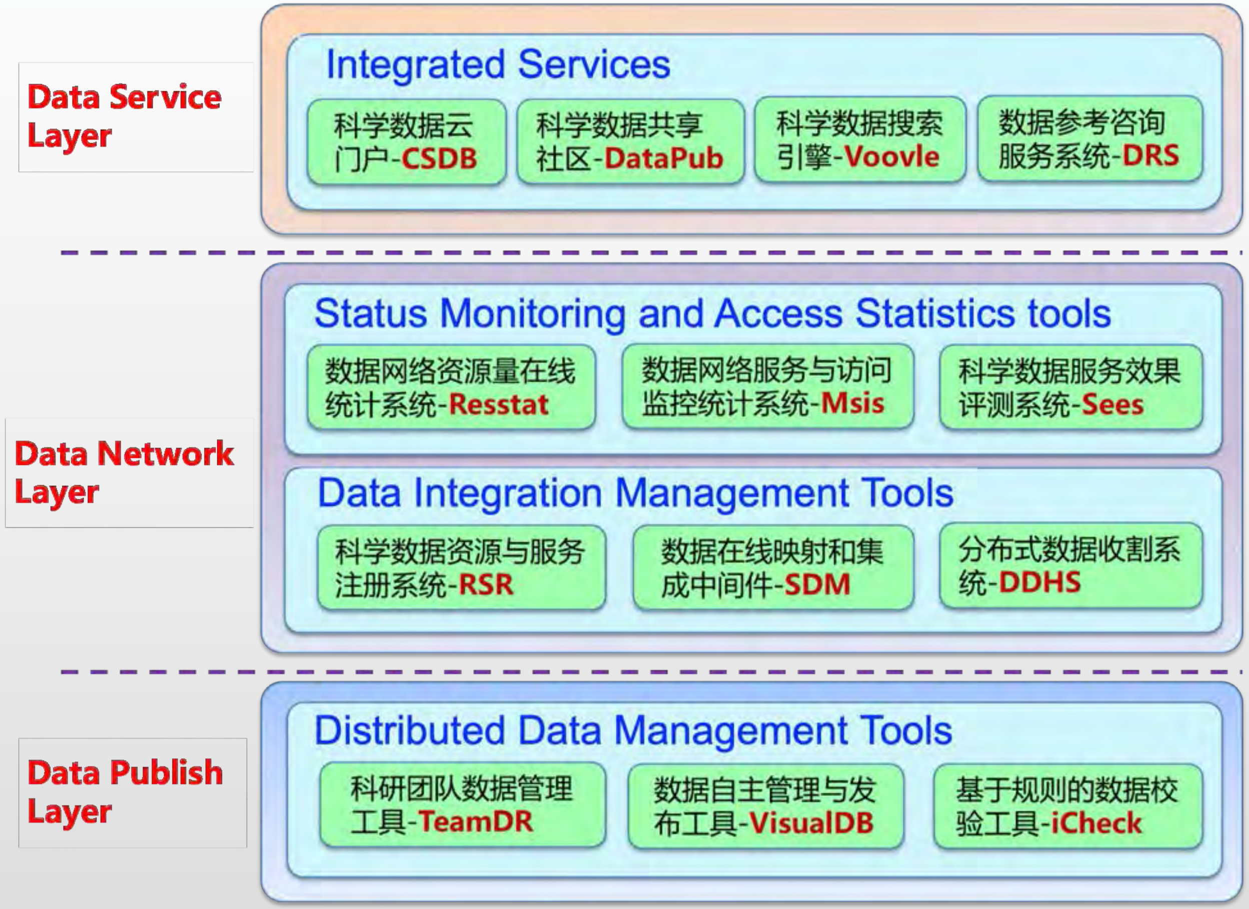
引种、保育、物候统计信息

序号	植物园名称	中文名称	拉丁名称	引种			保育			物候		
				科	属	种	科	属	种	科	属	种
1	华南植物园	被子植物门	Angiospermae	179	1389	6211	0	0	0	0	0	0
2	武汉植物园	被子植物门	Angiospermae	164	1037	6299	15[illegible]	748	4895	0	0	0
3	版纳植物园	被子植物门	Angiospermae	188	1771	7977	132	849	3219	0	0	0

总记录数：3 <<上一页 1 下一页>>

L4：Technical interoperability

- Scientific Data Softwares @ CAS project



Data Service Layer

Integrated Services

科学数据云门户-CSDB

科学数据共享社区-DataPub

科学数据搜索引擎-Voovle

数据参考咨询服务系统-DRS

Data Network Layer

Status Monitoring and Access Statistics tools

数据网络资源量在线统计系统-Resstat

数据网络服务与访问监控统计系统-Msis

科学数据服务效果评测系统-Sees

Data Integration Management Tools

科学数据资源与服务注册系统-RSR

数据在线映射和集成中间件-SDM

分布式数据收割系统-DDHS

Data Publish Layer

Distributed Data Management Tools

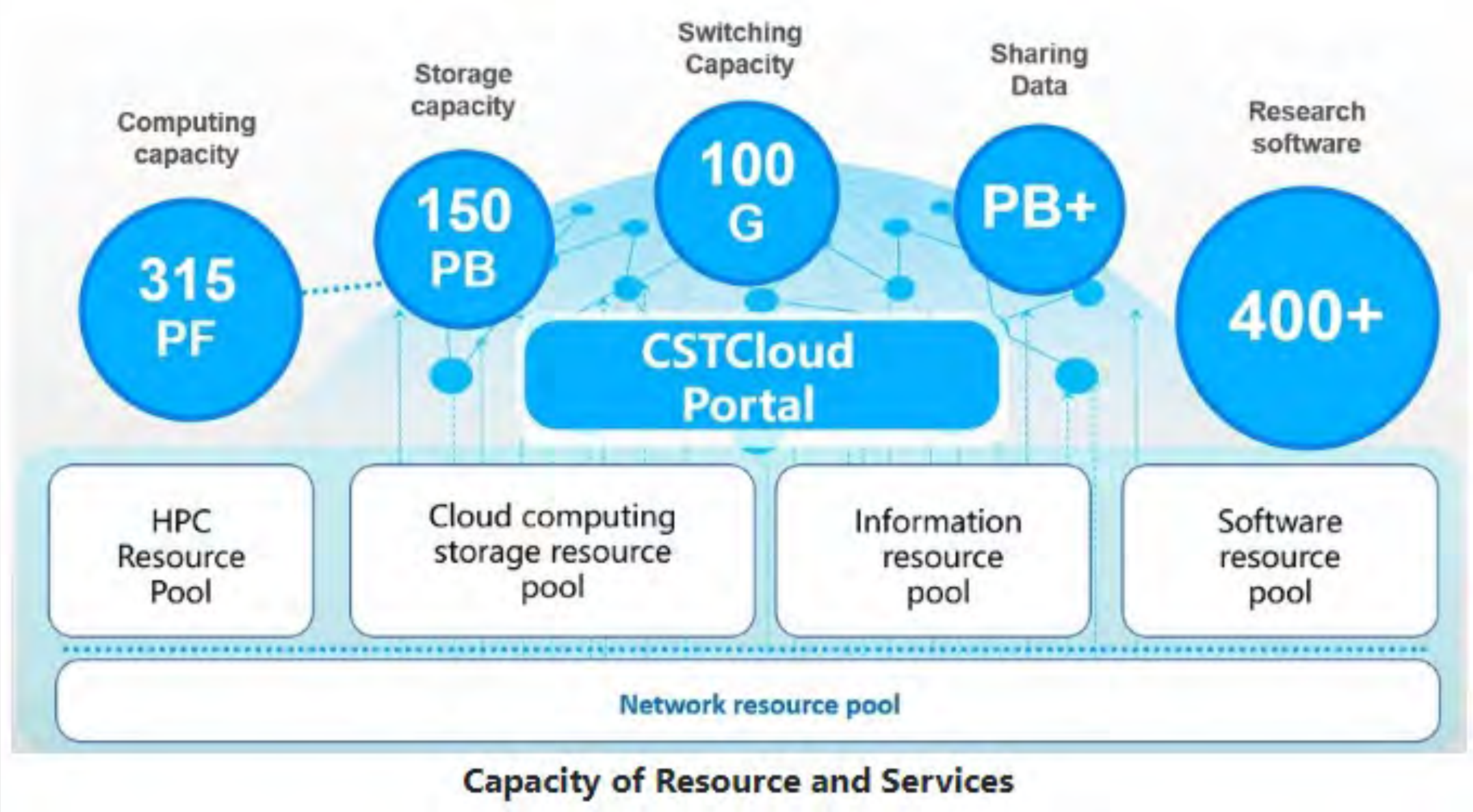
科研团队数据管理工具-TeamDR

数据自主管理与发布工具-VisualDB

基于规则的数据校验工具-iCheck

Technical interoperability -- cyber infrastructure

- China Science and Technology Cloud（CSTCloud）



- Network resource service
- Cloud storage service
- Object storage service
- Cloud backup service
- Computing service
- ......

→ GOSC

Contents

- Two projects
- DataIO practice
- Challenges

Challenges

- **The ecology of open and shared scientific data should be improved urgently**
  - The number of open and shared scientific data is still insufficient compared with the scale of data resources generated in China.
  - The availability and usability of scientific data that have been opened and shared are not high, the quality is uneven, and the overall level still needs to be improved.
  - The ecological environment for open and shared scientific data needs further improvement, including further policies and regulations, and exploration of innovative forms of open and shared data.

Challenges

- **The comprehensive utilization of scientific data is still far from enough**
  - The management mode of scientific data center is difficult to meet the high efficiency requirements of multiple applications.
  - The cross-disciplinary application is still in the demonstration stage.
- **The scientific data centers are lacking International influence**
  - Although many data centers have been formed, but their overall strength is not prominent, especially lacking international influence.
  - Data exchange and interoperation with international data centers are also severely missing

COMMITTEE ON DATA

CODATA

INTERNATIONAL SCIENCE COUNCIL

International Science Council

Thanks for listening