



# **Research on scientific data management in academic libraries**

Practice of Wuhan  
University Library

DingNing  
Wuhan University Library  
dingning@lib.whu.edu.cn

# CONTENTS



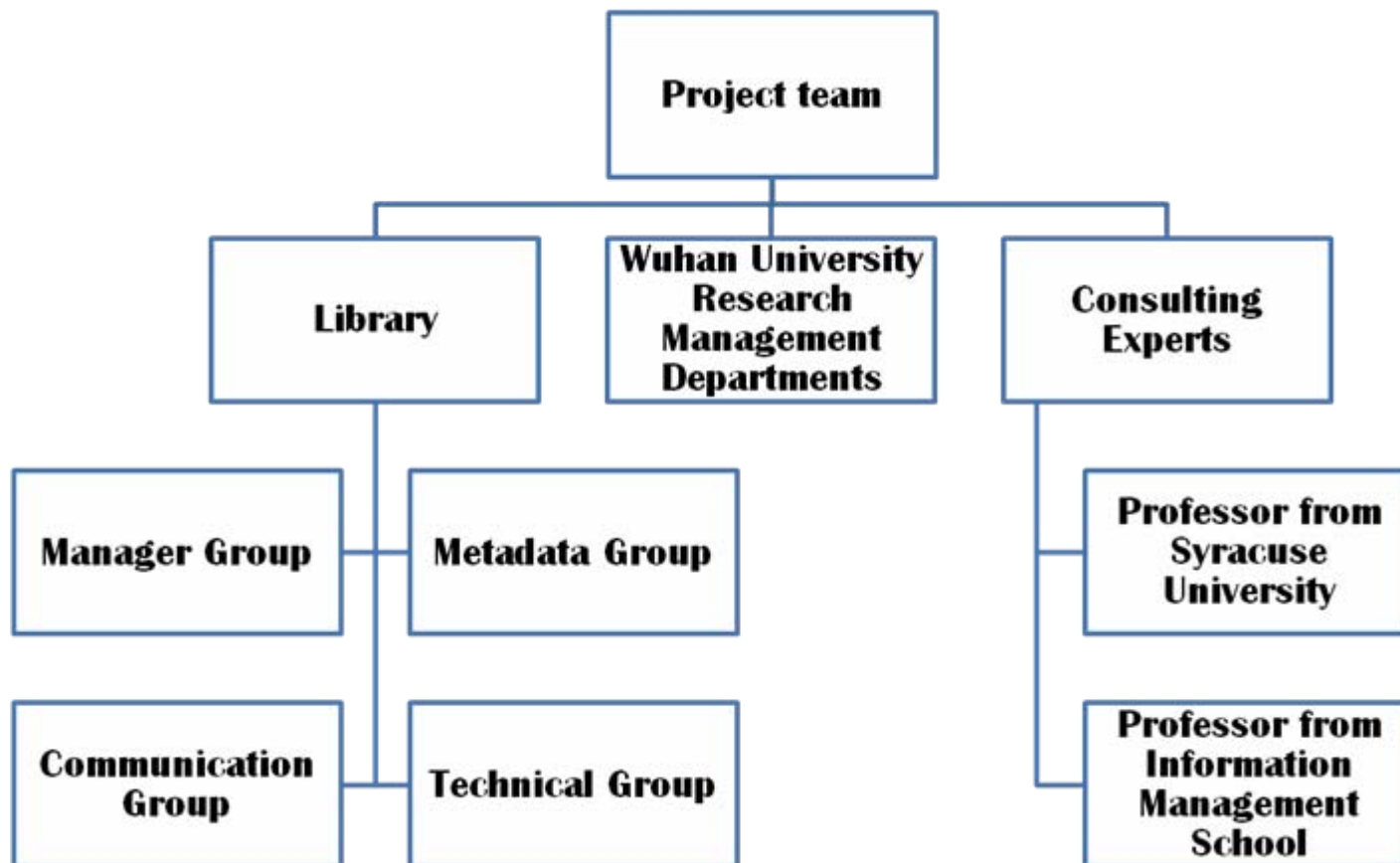


# BACKGROUND | WHY?

- Technology brings changes to scientific research
  - *E-Science, E-research, Big data*
  - *Data-sensitive, widely corporative*
- Data Management is becoming more and more important
  - *Many data sharing polices came on*
  - *A lot of scientific data sharing projects in the world*
- There isn't an interdisciplinary data sharing platform in Chinese University

# BACKGROUND

- A pilot project funded by CALIS
- Organizational structure



## Train of Thought

HOW?

- *Track the latest progress and trends of foreign scientific data management theory and practices*
- *Get the current situation of data management and scientific data literacy of researchers in Chinese Universities*
- *Choose 2-3 research institutions or subjects in Wuhan University as pilots and build a data management test platform*
- *Set a data service system and promote our platform and services*

## Current Situation at Home and Abroad | HOW?

- Methods: Literature survey + Webpage survey
- Data management projects in the world
  - *America, Canada, Australia, UK, Japan, China*
  - *Administrative mechanism, organizational structures*
- Data sharing platforms in the world
  - *Functions, software and technologies*
- Data management standards

## Needs of Researchers | HOW?

- Design a questionnaire contained 19 questions
  - *scientific data generation: 4 questions*
  - *scientific data sharing and management: 12 questions*
  - *Expectations and needs: 3 questions*
- Respondents:
  - *11 universities in Wuhan area*
  - *The faculties and postgraduates, 1:3*
- 1200 sending cents, 902 returned cents
- Time period: Nov.2011- Mar. 2012

- Data in university is in miniature, dispersive and sporadic
  - 69% data's magnitude is less than 1GB
  - 55% data is dispersed in postgraduates which have high mobility
  - The frequencies of data generation or data procurement of every team is not consistent
  - The source of data is quite different: lab experiment, field observation, social investigations, modeling or simulation, internet, purchase etc.



- Our researchers need more trainings on data management
  - *More than 60% researchers show that data has lost for not keeping them safely*
  - *More than 50% researchers never do permanent preservation to their data*
  - *73% researchers indicate that the background information of their data can't be access*
  - *50% researchers use their own standards to describe their data*
  - *Notion about data management is still lacking. 40% researchers are not satisfied or don't know data management can improve the release of achivements*

- The needs for data services and data management platform are strong
  - *Every option we provide is chosen in high percentage*
  - *30% researchers need library to develop a data management platform*
  - *30% researchers need library to provide data management consultation service*
  - *27% researchers think browsing, searching and downloading of data is most needed*

# Practice in Wuhan University



## PREPARATION

- Hold meeting with directors from “Academy of Humanities and Social Sciences” and “Academy of Science and Technology”
- Subject librarians contacted with researchers in almost every school
- Paid visits to *School of Water Resource and Hydropower Engineering* , *School of Life Science*, *School of Chemistry*
- Hold informal meeting with researchers in *Department of Sociology*

# PRINCIPLES & PILOTS

## PRINCIPLES

Larger quantity of data

Strong demanding to data management

Younger team

Disposed

Existing international or national data sharing platform

High Security requirement

Indisposed

## Pilots

- “Scorpion toxin project” in Life Science College
  - *Field observation data , protein/gene sequencing data*
- Department of Sociology
  - *Social investigation data*
- “Transmission dynamic of microblog project” in Information Management School
  - *Data caught from Internet automatically*
- Programs in Wuhan University Library
  - *Social investigation data*

# Workflow

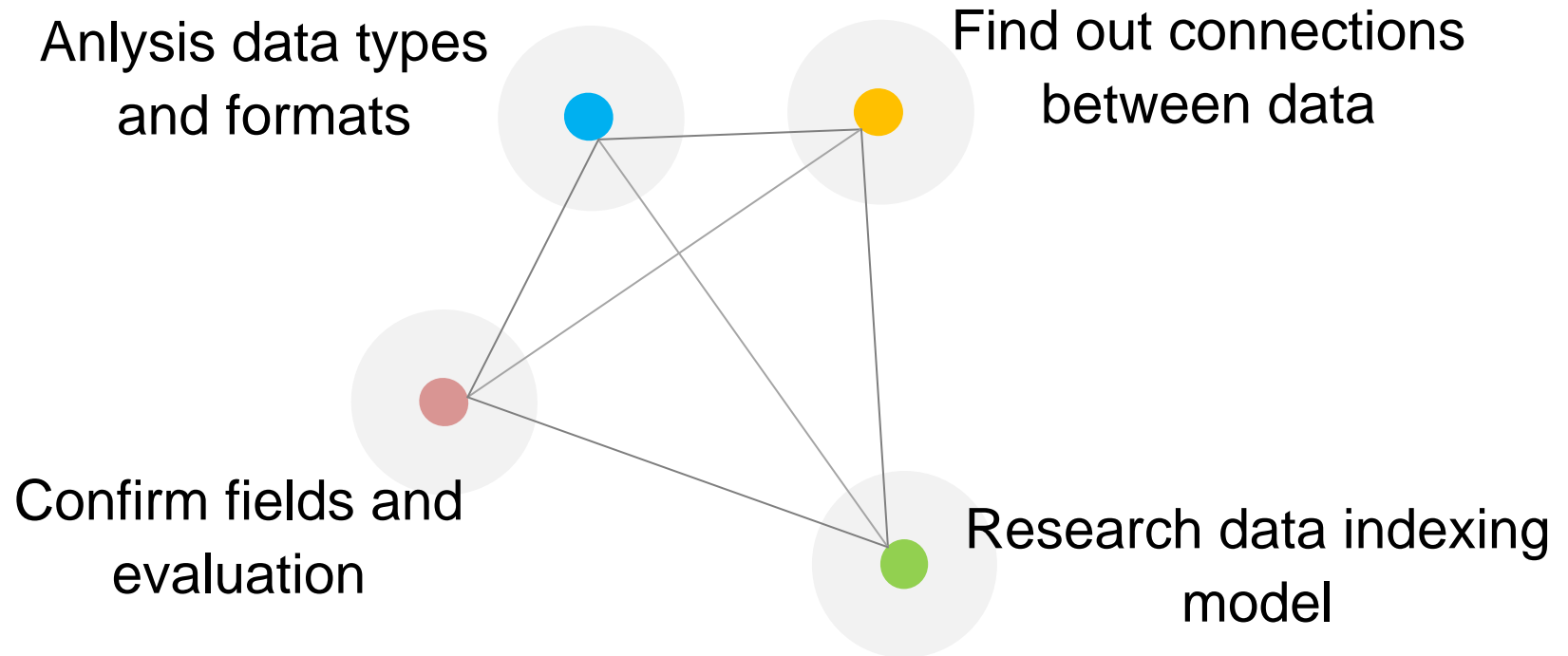
**1<sup>st</sup>** Subject librarians contact with researchers who have interests, and confirm corporation intention

**2<sup>nd</sup>** Get data samples and analysis the data structure, build metadata module

**3<sup>rd</sup>** Sustainly communication with researchers to do some redevelopements, website designs etc.

**4<sup>th</sup>** Researchers submit data and manage authority by themselves or entrust the work to library

# Metadata Design





# Metadata Design

## GenBank 字段详细说明和实例解析。

一级字段	二级字段	解释
Locus	Locus name	用于集合相似的序列，是唯一的。前三个字符，通常指定有机体；第四和第五个字符被用来显示其它组名称，如基因产品；对于分段的 entry，最后一个字母是连续整数系列之一。
	Sequence length	序列中核苷酸碱基对（氨基酸残基）的数量。
	Molecule Type	分子类型，可以包括的分子类型有：genomic DNA, genomic RNA, precursor RNA, mRNA (cDNA), ribosomal RNA, transfer RNA, small nuclear RNA 和 small cytoplasmic RNA。
	GenBank Division	记录所属的基因组数据库，用三个字母的缩写词表示。
	Modification Date	最后一次修改日期。

实例： LOCUS SCU49845 5028 bp DNA PLN 21-JUN-1999  
Locus 名称 序列长度 分子类型 GenBank 数据库 修改日期

位置 内容  
-----  
01-05 'LOCUS'  
06-12 space.  
13-28 Locus name.  
29-29 space.  
30-40 Length of sequence, right justified.  
41-41 space.  
42-43 bp.  
44-44 space.  
45-47 spaces, ss- (single-stranded), ds- (double-stranded), or, ms- (mixed-stranded).  
48-53 NA, DNA, RNA, tRNA (transfer RNA), rRNA (ribosomal RNA), mRNA (messenger RNA), sRNA (small nuclear RNA).  
Left justified.  
54-55 space.  
56-63 'linear' followed by two spaces, or 'circular'.  
64-64 space.  
65-67 The division code (see Section 3.3).  
68-68 space.  
69-79 Date, in the form dd-MMM-yyyy (e.g., 15-MAR-1991).

一级字段	二级字段	解释
DEFINITION		包括如源生物、基因名称/蛋白质名称或一些说明序列功能的信息。最后一行必须以句点结尾。定义字段的第二部分描述分子序列所示的基因和蛋白质信息。任何特别标识条款记载在中括号内。定义字段的第二部分记录在分隔方括号[]，提供有关的分子类型和长度的详细信息。

实例： DEFINITION Saccharomyces cerevisiae, TCP1-beta gene, partial cds, and Axl2p (AXL2) and Rev7p (REV7) genes, complete cds.

一级字段	二级字段	解释
ACCESSION		唯一标识符。该字段包含 6 位或 8 位的“流水号 (accession numbers)”。6 位字符的格式是：一个大写字母加上 5 位数字。8 位字符的则是两个大写字母加上 6 位数字。主号（第一个流水号）占用字符位 13-18/20。次号则以空格分隔，并存在多个次号的情况。

实例： ACCESSION U49845.  
ACCESSION AF181452.

一级字段	二级字段	解释
VERSION	GI	包含两种类型的标识符：复合的流水号 (accession number) 和 NCBI GI 标识符。复合流水号由固定的主号和按顺序增加的版本号构成，以句点分隔。NCBI GI 标识符则会随着变更重新分配。

实例： VERSION AF181452.1 GI601792.  
复合流水号 NCBI GI 标识符。

如果一个条目（例如，AF181452）有两个序列的变化，其 VERSION 字段的复合加入号的开始状态是 AF181452.1，经过一个变化后是 AF181452.2，再经过一个变化后是 AF181452.3。

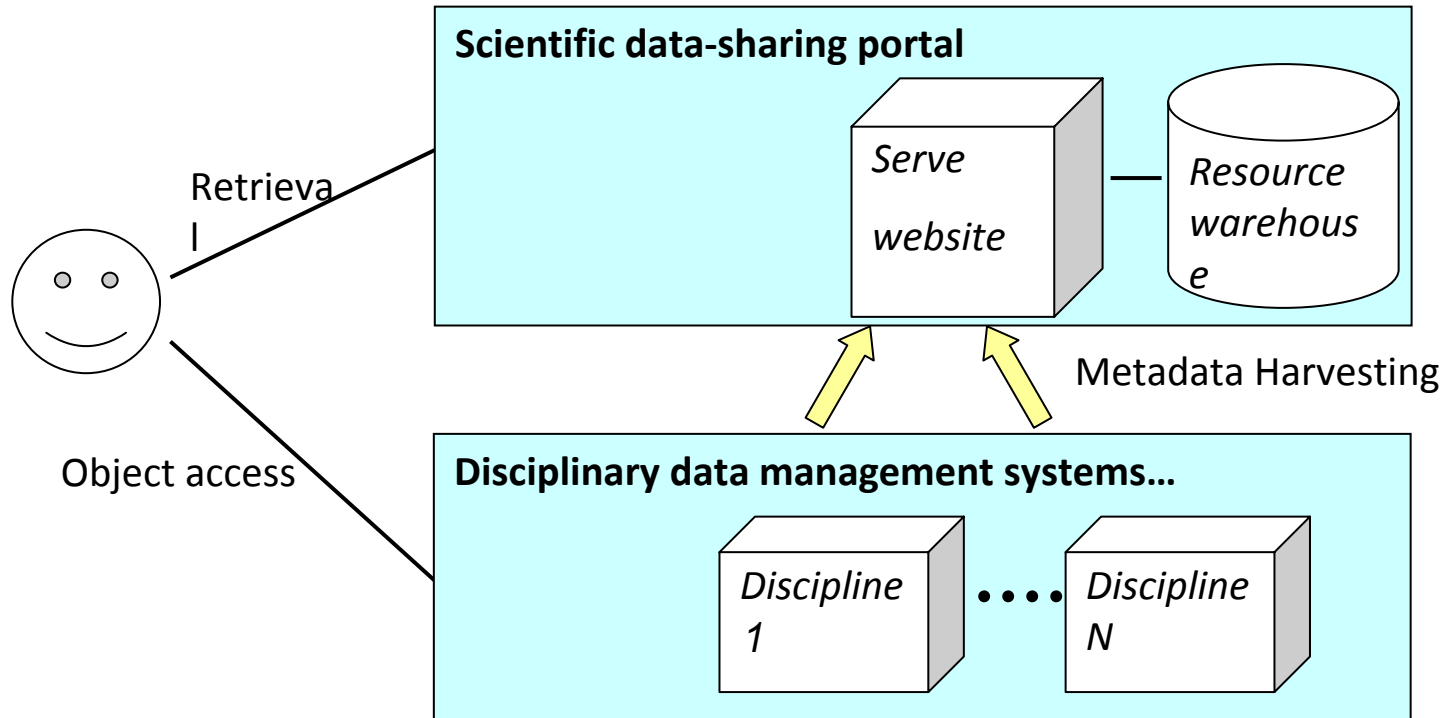
一级字段	二级字段	解释
KEYWORDS		关键字字段出现在需要注释的条目中，以分号分隔，以句点结束。在关键字字段为空的情况下，该字段仅包含一个句点。

一级字段	二级字段	解释
SOURCE	Organism	核酸或蛋白分子的来源物种。该字段包含两部分。第一部分是自由格式的信息，包括缩写形式的有机体名称和一个分子的类型。第二部分则以 ORGANISM 为字段名，包含来源有机体的科学名称、血统等信息，以分号分隔。

实例： SOURCE *Auicularia-judae* (mushroom) ribosomal RNA.  
ORGANISM *Auicularia auricula-judae*.  
Eukaryota; Fungi; Eumycota; Basidiomycotina; Phragmobasidiomycetes;  
Heterobasidiomycetidae; Auriculariales; Auriculariaceae.

一级字段	二级字段	解释
REFERENCE		REFERENCE 包含参考文献编号和（在括号内）该文献涉及的序列碱基范围。
	AUTHORS	按照在文中出现的顺序索引，以句点结尾。

# System structure



**1**  
Gate: one-stop  
portal

**2**  
Stages: sharing  
portal & disciplinary  
systems

**3**  
Levels: digital items-  
data sets-  
colleges/institutions



# System Development

HOW?

- Use open-source software Dspace and Java to build data management systems
- Use Lucene as system search engines and implement Chinese information retrieval by using Lucene Chinese word segmentation system or others
- Deploy parameters: *fields, titles, display*
- Do some redevelopment of DSpace
  - *Optimization to Information Retrieval*
  - *Chinese localization*
  - *Change data display*
  - *Increase customize functions: BLAST*

# China Academic Scientific data service





## 高校科学数据共享平台

China Academic Scientific Data Service

[登录](#)

快速检索

[搜索](#)

### 关于我们

本平台由教育部“211工程”三期建设支持，是中国高等教育文献保障系统（CALIS）项目下的一个预研类项目。项目由武汉大学图书馆主持，旨在研究科学数据平台建设流程、方法及科学数据管理的各类标准规范。以用户为中心，了解试点研究项目对于科学数据平台建设方法及科学数据管理与图书馆员合作方式的具体要求；以需求为驱动，实施试点项目的科学数据管理；结合文献调研结果与试点项目的经验.....

[更多内容>>](#)

### 数据浏览



#### 武汉大学生命科学学院

[> MORE](#)

- 蝎物种资源数据库
- 蝎物种遗传基因资源数据库
- 蝎物种遗传蛋白资源数据库



#### 武汉大学社会学系

[> MORE](#)

- 武汉市远城区城镇化综合评价指标体系研究
- 用人单位对武汉大学毕业生的全面评价研究



#### 武汉大学信息管理学院

[> MORE](#)

### 推荐网站

- DataONE
- Data Conservancy
- Data Curation

# Scorpion space



[Home](#) [Browse](#) [Search](#) [Blast](#) [About](#)

## Quick Search

  
[Advanced Search](#)

## Quick Links

- [Scorpion Nucleotide](#)
- [Scorpion Protein](#)
- [Scorpion Species](#)

## Related Links

- [NCBI](#)
- [Scorpion files](#)
- [The Scorpion Fauna](#)
- [Arachnodata](#)
- [The Spiral Burrow](#)
- [Patrick's Scorpion Page](#)
- [Kari's Scorpion Pages](#)

[SDM Home](#) >

## Species and Toxins

In:

Scorpion is an ancient venomous animal. The first scorpion is believed to have evolved from the Eurypteridae or water scorpions 425 to 450 million years ago in the middle of the Silurian epoch. During long natural evolution, about 1500 species of scorpions are known to mankind, and about 25 are capable of causing human death. Meanwhile, scorpion can produce various toxins, which have been proved to be valuable tools or drug leads in both basic and medical researches. SSTB is a scorpion resource database, which contains scorpion species in China and toxins in the world. If you encounter any problem or have any suggestions, please don't hesitate to contact us by E-mail ([liwxlab@whu.edu.cn](mailto:liwxlab@whu.edu.cn)).



## Collections in this community

[Scorpion Nucleotide](#)

[Scorpion Protein](#)

[Scorpion Species](#)

## Recent Submissions

[Buthus martensii toxin BmP01 precursor \(BmP01\) gene, complete cds](#)

[1PE4\\_A 67 aa linear INV 10-JUL-2009](#)

[Buthus martensii anti-neuroexcitation peptide II precursor \(ANEPII\)](#)

[Buthus martensii putative sodium channel toxin BmKT mRNA, complete](#)

[Buthus martensii toxin TXKs4 mRNA, complete cds.](#)



# Functions

- Data submission/Preservation
  - *Researchers submit their data*
  - *Library keep their data safe and accessible*







# Functions

- Browse by Category
  - *college/institution, project/dataset, title*
  - *Metadata/details*

The screenshot displays the Scorpion Space website. The header features the 'Scorpion space' logo with a red scorpion icon and navigation links: Home, Browse, Search, Blast, and About. On the left sidebar, there is a 'Quick Search' box with a 'Go' button and a link to 'Advanced Search'. Below this are 'Quick Links' for Scorpion Nucleotide, Scorpion Protein, and Scorpion Species, followed by 'Related Links' including NCBI, Scorpion files, The Scorpion Fauna, Arachnodata, The Spiral Burrow, Patrick's Scorpion Page, and Kari's Scorpion Pages. The main content area is titled 'Browsing "Species and Toxins" by Title' and includes a 'Jump to' section with a list of letters (0-9 A-Z) and a text input field for 'or enter first few letters:'. Below this, it shows 'Showing results 1 to 10 of 81' with a 'next >' link. The results list includes entries for '1PE4\_A 67 aa linear INV 10-JUL-2009' (Centruroides noxius) and 'AF095781 361 bp DNA linear INV 19-OCT-1999' (Smith, Donald Jr), both with links to 'Scorpion Nucleotide'. The bottom of the page shows a Windows taskbar with the Internet Explorer icon and a 100% zoom level.

**Scorpion space**

Home Browse Search Blast About

**Quick Search**

[Advanced Search](#)

**Quick Links**

- [Scorpion Nucleotide](#)
- [Scorpion Protein](#)
- [Scorpion Species](#)

**Related Links**

- [NCBI](#)
- [Scorpion files](#)
- [The Scorpion Fauna](#)
- [Arachnodata](#)
- [The Spiral Burrow](#)
- [Patrick's Scorpion Page](#)
- [Kari's Scorpion Pages](#)

[SDM Home >](#)

**Browsing "Species and Toxins" by Title**

Jump to: [0](#)[1](#)[2](#)[3](#)[4](#)[5](#)[6](#)[7](#)[8](#)[9](#)[A](#)[B](#)[C](#)[D](#)[E](#)[F](#)[G](#)[H](#)[I](#)[J](#)[K](#)[L](#)[M](#)[N](#)[O](#)[P](#)[Q](#)[R](#)[S](#)[T](#)[U](#)[V](#)[W](#)[X](#)[Y](#)[Z](#)

or enter first few letters:

Showing results 1 to 10 of 81  
[next >](#)

[1PE4\\_A 67 aa linear INV 10-JUL-2009](#)  
[1PE4\\_A 67 aa linear INV 10-JUL-2009](#)  
*Centruroides noxius* (Mexican scorpion)  
[Scorpion Nucleotide](#)

[1PE4\\_A 67 aa linear INV 10-JUL-2009](#)  
[1PE4\\_A 67 aa linear INV 10-JUL-2009](#)  
*Centruroides noxius* (Mexican scorpion)  
[Scorpion Nucleotide](#)

[AF095781 361 bp DNA linear INV 19-OCT-1999](#)  
*Smith, Donald Jr*  
[Scorpion Nucleotide](#)

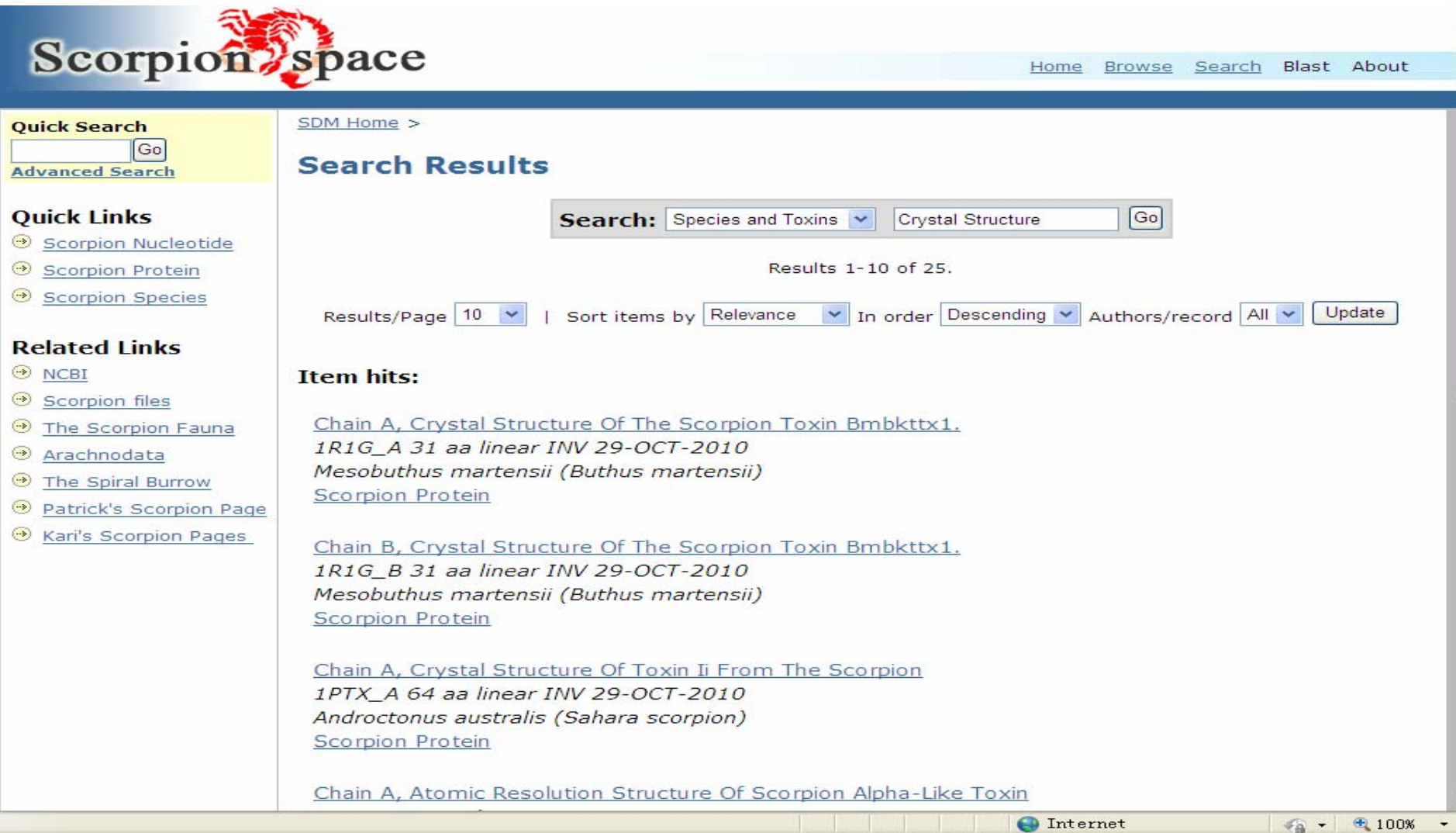
[Buthus martensii alpha toxin 1 precursor, mRNA, complete cds.](#)  
[AF288607 408 bp mRNA linear INV 21-MAY-2002](#)  
*Mesobuthus martensii* (*Buthus martensii*)  
[Scorpion Nucleotide](#)

五笔型 ● " "

Internet 100%

# Functions

- Retrieval in every fields



The screenshot displays the Scorpion Space website interface. The header features the 'Scorpion space' logo and navigation links: Home, Browse, Search, Blast, and About. A left sidebar contains sections for 'Quick Search' (with a search box and 'Go' button), 'Quick Links' (with links to Scorpion Nucleotide, Scorpion Protein, and Scorpion Species), and 'Related Links' (with links to NCBI, Scorpion files, The Scorpion Fauna, Arachnodata, The Spiral Burrow, Patrick's Scorpion Page, and Kari's Scorpion Pages). The main content area shows 'Search Results' for the query 'Crystal Structure'. It includes a secondary search bar, a result count of 'Results 1-10 of 25.', and filters for 'Results/Page' (10), 'Sort items by' (Relevance), 'In order' (Descending), and 'Authors/record' (All). The 'Item hits' section lists three results, each with a title, accession number, date, species, and a link to the 'Scorpion Protein' page.

**Scorpion space**

Home Browse Search Blast About

Quick Search  
[Search Box] Go  
Advanced Search

Quick Links  
Scorpion Nucleotide  
Scorpion Protein  
Scorpion Species

Related Links  
NCBI  
Scorpion files  
The Scorpion Fauna  
Arachnodata  
The Spiral Burrow  
Patrick's Scorpion Page  
Kari's Scorpion Pages

SDM Home >

**Search Results**

Search: Species and Toxins Crystal Structure Go

Results 1-10 of 25.

Results/Page 10 | Sort items by Relevance In order Descending Authors/record All Update

**Item hits:**

[Chain A, Crystal Structure Of The Scorpion Toxin Bmbkttx1.](#)  
1R1G\_A 31 aa linear INV 29-OCT-2010  
*Mesobuthus martensii* (*Buthus martensii*)  
[Scorpion Protein](#)

[Chain B, Crystal Structure Of The Scorpion Toxin Bmbkttx1.](#)  
1R1G\_B 31 aa linear INV 29-OCT-2010  
*Mesobuthus martensii* (*Buthus martensii*)  
[Scorpion Protein](#)

[Chain A, Crystal Structure Of Toxin Ii From The Scorpion](#)  
1PTX\_A 64 aa linear INV 29-OCT-2010  
*Androctonus australis* (*Sahara scorpion*)  
[Scorpion Protein](#)

[Chain A, Atomic Resolution Structure Of Scorpion Alpha-Like Toxin](#)

Internet 100%





# Functions

- Download records

**抽样检验方法：** 问卷采取多阶段抽样的方法，根据每个区的行政单位的不同情况，从中以街道为抽样单位，再从随机抽中的街道中抽取样本，组成一个容量为300的样本，每一个居民被抽中的概率是均等的，保证了样本的随机性。问卷采取无记名的形式，每个区发放50份问卷，共300份问卷，回收有效问卷297份，有效率达到99%。

**空间覆盖范围：** 武汉市各远城区

**时间覆盖范围：** 2011年7月 - 2011年11月

样本  
发放问卷的  
回收问卷的  
有效样本的  
有效问卷回  
数据  
所属



<a href="#">czh.doc</a>	调查分析报告	2.79 MB	Microsoft Word	<a href="#">浏览/打开</a>
<a href="#">czhdctjb.xlsx</a>	调查统计表	115.3 kB	Microsoft Excel	<a href="#">浏览/打开</a>



# Functions

- Patron management
  - *Devise passwords*
  - *Add new users*
- Authority management
  - *Browse metadata, browse data/datasets, download metadata, download data/datasets*
- Distribution management
  - *Open Access*
  - *Accessed by special groups*



# Functions

[SDM Home](#) >  
[general.administer](#) >

## Administer Authorization Policies

Choose a resource to manage policies for:

Manage a Community's Policies

Manage Collection's Policies

Manage An Item's Policies

Advanced/Item Wildcard Policy Admin Tool

<b>Collection:</b>	反剽窃实现下的相似信息动力传播学研究 图书馆读者调查数据 武汉市远城区城镇化综合评价指标体系研究 用人单位对武汉大学毕业生的全面评价研究
<b>Content Type:</b>	item
<b>Group:</b>	<div>Administrator Anonymous COLLECTION_23_SUBMIT COLLECTION_3_SUBMIT COLLECTION_3_WORKFLOW_STEP_3 COLLECTION_69_SUBMIT COLLECTION_70_SUBMIT COMMUNITY_1_ADMIN COMMUNITY_24_ADMIN COMMUNITY_29_ADMIN</div>
<b>Action:</b>	READ

Add Policy

Clear Policies

(warning: clears all policies for a given set of objects)

# Functions

[高校科学数据共享平台](#) >

[武汉大学图书馆](#) >

快速链接

[图主](#)

[图主](#)

**调查实施者:** 胡永生、刘兵红、柳卫莉、赵基明

**访问员:** 胡永生、刘兵红、柳卫莉、赵基明

**审核员:** 胡永生

**作者/存档者:** [胡永生](#)

[新使用者请点击此注册。](#)

请在以下字段中输入您的e-mail与密码。

e-mail:

密码:

登录

[忘记密码?](#)

发回

**有效样本的数量:** 326

**有效问卷回收率:** 86%

**数据状态:** 已完成

**版本:** 1.0

**所属集合:** [图书馆读者调查数据](#)

计划

文件中的档案:

档案	描述	大小	格式	
<a href="#">kwj.doc</a>	理科学科服务调查表	47.5 kB	Microsoft Word	<a href="#">浏览/打开</a>



# PROMOTION

- Held two informal meeting with research administrative secretaries of 37 schools/departments and a few researchers/research teams
  - *Professor QinJian introduced progress of data sharing and data management at abroad*
  - *Introduction on our data sharing platform*
  - *Feedback from secretaries and researchers*



# EXPERIENCES & PROBLEMS

FUTURE

- On the view of microscopic, there are some key issues in data management
  - *Library participates in data management work, but may not direct contact with data itself*
  - *Scientific data management should emphasize access controls and availability*
  - *It is important to find out different requirements of researchers and provide services according to what they exactly need*



# EXPERIENCES & PROBLEMS

FUTURE

- **Problems 1:** Even that we corporate with research management departments, the promotion effect is not that ideal.
  - *Promotion makes some progresses: some researchers showed their interests but the work still delayed for a lot reasons.*
  - *Our work is in initial starting, we need force power to push it*



## EXPERIENCES & PROBLEMS

| FUTURE

- On the view of macroscopic, scientific data management in Chinese universities requires policy support and constructing mechanism
  - *No national foundation policies come into being*
  - *In university, the corporation amount different departments, just like library, research management departments, IT department is really important.*
  - *University administrations should make policies and show responsibility for scientific data management work in the university*





# EXPERIENCES & PROBLEMS

FUTURE

- **Problem 2:** The demands of researchers are changing, increasing. New questions are raised from time to time.
  - *Big data submission and transformation*
  - *Dspace can't satisfy some data formats, like sequence data*
  - *Permanent preservation*
  - *Data back-up*



# EXPERIENCES & PROBLEMS

FUTURE

- Our system must improve continuously to satisfy the changing demands of researchers
  - *Semiautomatic data indexing*
  - *Automatic or semiautomatic transformation in different platform*
  - *Customize functions*
  - *System stability, security, durative*



# EXPERIENCES & PROBLEMS

| FUTURE

- **Problem 3:** Data service is one of personalize services and it requires a lot of human input. Here comes a paradox between data service and routine.
  - *Every discipline data system is a small project and the whole project team will involve*
  - *What should we plan in the future?*



## EXPERIENCES & PROBLEMS

| FUTURE

- We should divide our target into some subgoals and make it clear that what we can do at every stage
  - *At present, our goal is only to set up a platform and to provide basic services*
  - *In the future, we should gradually increase the quantity of pilots and data, and improve the construction of mechanism of Wuhan University, and provide some advanced services*



# Thank You!



DingNing  
Wuhan University Library  
[dingning@lib.whu.edu.cn](mailto:dingning@lib.whu.edu.cn)

