

Construction Scheme of NSFC Open Access Library

Jianjun Li¹, Minshe Zhang¹, Dong Li¹, Wei Zhang¹ and Jin Wang^{1,2}

¹ *Information Center, National Natural Science Foundation of China, Beijing
China*

² *College of Information Engineering, Yangzhou University, Yangzhou, China*

Abstract

With the fast development of Open Access (OA) in recent years, it helps stimulate free scientific achievements propagation via Internet and promote academic exchange and fast publishing in an efficient and cheap way. In this paper, we propose a construction scheme of our Natural Science Foundation of China (NSFC) open access library which is under development recently. We first propose our overall architecture for OA library. Then, we present detailed design from upper layer OA webpage with 3-level display hierarchy to the middle layer some key function modules. Next, bottom layer raw data acquisition module and other important system function modules like interface module, people management and security modules are explained with illustrative figures and table.

Keywords: *Open Access (OA), NSFC, OA Journal, OA Repository, E-print*

1. Introduction

Open Access (OA) has gained fast development in recent years due to strong support from international academic, publishing and library information area. The initiative purpose of OA is to stimulate free scientific achievements propagation via Internet, to promote academic exchange and publishing, to improve the utilization degree of research achievements and to ensure lasting preservation of scientific information [1].

Based on the explanation from Budapest Open Access Initiative (BOAI) in 2002, any user can read, download, copy, distribute, print, search and link to the full texts of OA documents, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose [2]. Usually, there are two ways to achieve open access: 1) to publish articles on some OA journals; 2) to self-archive articles in some OA repositories [3]. Until 2010, the Directory of Open Access Journal (DOAJ) has collected 4953 OA journals, and the OA repository number has reached 1620 inside the OpenDOAR (Directory of Open Access Repositories).

Natural Science Foundation of China (NSFC) is a governmental organization which is in charge of managing national natural science projects. Its main tasks include supporting basic research, fostering talented researchers, developing international cooperation and promoting socioeconomic development [4].

Founded in 1986, NSFC has accumulated huge amount of project achievements and management data in the past 30 years, ranging from projects, person, affiliation, achievements (*e.g.* papers, technical reports, patents, dissertations and awards etc.) and various project reports (*e.g.* progress report, mid-term report, final report etc.). Based on some statistical reports, there are about 1 million project application proposals, 0.3 million granted projects, 8 million participants, 5 million achievement information, 3000 affiliations and 1.8 million electronic items inside the NSFC data repository [5].

NSFC released an OA policy statement for research publications of its funded projects on May 15 2014 formally [6]. As a member of Global Research Council (GRC), NSFC basic knowledge library collect full paper as achievement from relevant projects which

are supported by NSFC. The main purpose of NSFC OA library is to provide open access to the public researchers, to propagate state-of-the-art scientific results from basic research area, and to promote advancement of science and technology. According on this policy, NSFC OA Library is under development at the first stage based on the rich project relevant academic achievements of NSFC. Main reasons to construct OA library include three aspects. First, the NSFC OA library can be beneficial to those researchers for free and convenient open access. Second, it can help NSFC better manage projects and relevant achievements. Finally, a benign academic bio-system can be built via activities like paper upload/download, scientific research exchange and collaboration etc.

The rest of this paper is organized as follows. Some related work about OA is first presented in Section 2. Then, the overall OA system architecture is given in Section 3. Detailed design from webpage design module (upper layer), main functional modules (middle layer), data acquisition module (bottom layer) etc. are provided in Section 4. Section 5 concludes this paper.

2. Related Work

The idea of free access to online articles arose long time ago before the term OA was formally proposed. Many researchers had already self-archived their academic documents in certain FTP for free access since 1970s, while the formal OA concept was first presented in 2002 BOAI, as mentioned above.

OA online digital resource is a novel and important academic information resource, which provide users free and convenient access to some full text journal papers. Compared with the old fashion publishing patterns, it is a novel pattern of invention and milestone for publishing area. In many scientific disciplines, OA articles are of more academic research value than the non-OA ones from the aspect of convenient availability for reference.

The primary two ways to gain OA are OA journal and OA repository, as is stated above.

For OA journal which is also called “golden OA”, it provides user from all over the world peer-reviewed, online, free articles on its own journal database. There is no restriction on price or copyright issues. In case when OA journals charge processing fees, it is the author's employer or some funder who usually pays that fee instead of the individual author.

For OA repository which is also called “green OA”, it saves articles in certain university or institutional repository. Such articles may include draft, revised or final version of their paper, or scanned hardcopy of published ones. Some lecture notes or ppt slides can also be saved there. Compared with OA journal, it provides user more convenient and faster access to journals, while the articles quality is not as strictly controlled as OA journals which are peer-reviewed and finally published. Thus, authors need to discriminate articles quality by their own judgment. In fact, the idea of self-archiving documents had already been practiced by computer scientists on FTP since 1980s.

A third way to gain OA is e-print articles, which have not yet been formally published. E-print articles also have very high academic value since they help promote academic exchange and resource sharing. They also help researchers track the latest research progress and avoid repeated work, since it takes much time to get the finally published formal journal articles. The authors will voluntarily submit their own e-print articles to some database or directory, and they take full responsibility for their behavior. The merits of e-print articles mainly include fast exchange, interactive, resource sharing etc.

Another important way to gain OA articles is via authors' personal homepage. With permission from publisher, authors can put their own works on the homepage for propagation or exchange if there might be some copyright problem. They can also put

their draft version, technical report or revised version on homepage, which is usually free from copyright debate.

Table 1 lists some famous and representative OA journal and OA repository sites, which are of much academic value for reference.

Table 1. Some Representative OA Journals and Repositories

Three Ways	Entry	Description
OA Journal	http://cnplinker.cnpeak.com [7]	Open journal link service platform developed by Chinese publishing group
	http://www.doaj.org/ [8]	DOAJ provides peer-reviewed high quality journals
	http://www.highwire.org/lists/freeart.dtl [9]	Globally the largest free access publishing agency
	http://www.ncbi.nlm.nih.gov/pmc/about/openftlist.html [10]	PMC OA mainly in life science research area
	https://www.openj-gate.com/ [11]	Open J-Gate
	https://www.jstage.jst.go.jp/browse/-char/en [12]	Leading OA journal library in Japan
OA Repository	http://www.dspace.org/ [13]	MIT and HP jointly developed
	http://repository.ust.hk/dspace/ [14]	HKUST developed repository using DSpace
	http://eprints.anu.edu.au/ [15]	Australian National University OA repository
	http://libguides.caltech.edu/CODA [16]	Caltech Collection of Open Digital Archives (CODA)
E-print articles	http://arxiv.org [17]	More than 1 million e-prints maintained by Cornell Univ.
	http://www.paper.edu.cn/ [18]	Science paper online library

Developed by the China National Publications Import and Export (Group) Corporation in 2002, cnpLINKer platform provides about 30000 journals information from 3000 journal publishing companies covering 50 countries or districts among which some of them are free. More than 20 millions of full text link resource and abstract information are provided.

Directory of Open Access Journals (DOAJ) is viewed as an online OA directory system which provides peer-reviewed high quality journals, which are published simultaneously with formal journals with free access. It is developed and maintained by Lund University library in Sweden. It includes 5690 journals information, where 2436 journals are open accessible with about 0.5 millions journal articles for free.

HighWire Press is the globally largest academic publishing agency which was found in 1995 by Stanford University library. Now, it has more than 2 millions free text articles and 7 millions journals all together. It covers research areas ranging like life science, medicine, physics and social science.

PubMed Central (PMC) was developed by National Center for Biotechnology Information (NCBI) which is affiliated with NIH in 2000. It mainly covers about 2000 life science relevant journal articles, where 150 journals are open accessible. The data source is from Blackwell Online Open and Springer Open Choice.

Open J-Gate was developed by Informatics (India) Ltd in 2006. It includes about 6000 journals where more than 4000 journals are strictly peer-reviewed. It provides three search methods, which are quick search, advanced search and browse by journals. Full text journal link information is provided and timely updated.

Japan Science and Technology Information Aggregator, Electronic (J-STAGE) is a leading OA journal library in Japan which includes 1737 journals, 127 proceedings with more than 2.6 millions articles.

OA depository can save not only pre-print version but also post-print version articles. In accordance with the pre-print version, post-print version articles are those which are finally published open after strict peer-reviewed procedure. There are two kinds of OA depository libraries, namely institution based and discipline based library.

DSpace system was jointly developed by MIT library and HP Company in 2002. It is an open source software platform written in Java with UNIX system, which is widely used by universities and institutes worldwide. The electronic document formats it supports include articles and preprints, technical reports, conference papers, datasets, images, audio/video documents etc.

Hong Kong University of Science and Technology (HKUST) developed an OA repository using DSpace software since 2004, which saves about 64,885 records, 8040 documents and 537 scholar profiles submitted by HKUST faculty members. Such documents and records can be sorted via institute, department, community or via author, key word etc.

Some other famous OA repositories include Australian National University OA repository, Caltech Collection of Open Digital Archives (CODA) etc. which cover online open journals, TR, conferences, book chapters etc.

E-print arXiv was developed by the Los Alamos national lab in 1991 with the support from NSF and DoT (Department of Energy). From 2001, this e-print library is maintained by Cornell University library. It provides more than 1 million e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics areas. Document formats vary from PS, pdf to DVI etc.

Science paper online library was developed by Chinese Ministry of Education in 2006. Its main purpose is to help high quality quick publication in order to promote academic exchange. Traditional procedures like paper review, modify, re-editing and publishing are shortened to a large degree therein.

3. Overall Architecture

The overall architecture of NSFC OA library is shown in Figure 1. We will briefly mention some features of each module while details will be explained in the next section.

The upper layer OA webpage is an interface between system and users. OA webpage has three-level display hierarchy. On the first level, users can search their interested contents by typing some key words like author, affiliation, journal title etc. On the second level, some searched results will be displayed based on certain rules like similarity degree, year or citation number. If users are interested about certain journal, they can go to the third level to see detailed information about that journal and download it.

On the middle layer, some key function modules are necessary such as data searching, data collection, data supplement and data clearing modules. Also, system should provide users some policy reminding function to prevent misuse of data, and the authors can upload or claim their own academic achievements after logging in the system. Management module is also very important since all the data need to be managed in an efficient way, such as journals, patents, awards and dissertations etc. Different types of data need to be treated differently like e-print, modified version or final version journals, ppt files and audio/video files.

On the middle left part is system interface module, which provides an interface between final clean data and other modules like OA webpage, security and people modules. After data clearing, collection and supplement, data are viewed as clean data rather than filthy data which can be used for safe later on. On the middle right part are people management and security modules respectively. System should discriminate

different role of users like system administrator, expert, normal user and guest. Detailed people relevant information needs to be carefully organized and managed. Security module is an important part for the system which provides access control, authentication, prevention from attack functions etc.

On the bottom layer is the data acquisition module. Data need to be collected from various sources in a periodic manner with update. Usually, such raw data are called “filthy data” since they are usually incomplete and even erroneous. Further processing is needed such as data clearing, supplement and collection etc.

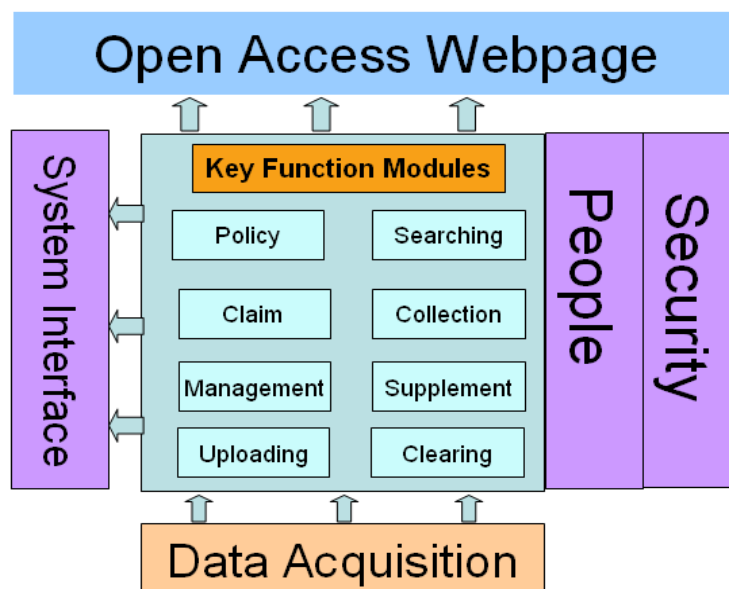


Figure 1. Overall Architecture of OA Library

4. Detailed Design of OA Library

4.1 Open Access Webpage

The OA webpage is a three-level display hierarchy, as is mentioned above. Figure 2 shows the first layer display content which mainly include NSFC logo, searching toolbar, OA policy statement, user guide, login button, Chinese/English version and hot paper ranking etc.

It is worth mentioning that there are more than 1 million electronic documents in the NSFC database depository for the past 30 years. The main purpose of OA library is to stimulate academic exchange by providing scientists an OA platform to access their interested articles in a timely and efficient way to facilitate their research work. NSFC supports basic natural science research from the following 8 disciplines, namely mathematics and physics, chemistry, life science, earth science, engineering and material science, information science, management science and medicine. For example, there are about 23289 journal articles in the life science discipline area from our OA library, as is shown in Figure 2.

Other functions like navigation toolbar, useful links, contact us, website map etc. are also provided with 7-24 technical system support.



Figure 2. Layout of OA Webpage: the First Level

Figure 3 shows the searching results once a user types certain keyword like “reflection”. In such case, about 22 results are listed on the second level based on certain rules like similarity degree, title or year. We can also see from the left part that there are about 7219 articles in the mathematics and physics discipline and 6112 articles in life science discipline.

Figure 4 shows the third level searching results if a user is interested about certain journal article content. On the third level webpage, detailed information such as title, author list, affiliation, journal name, relevant granted project number, discipline area, DOI, year and link information are all listed. Most importantly, a download icon is provided for immediate access of that article.



Figure 3. Layout of OA Webpage: the Second Level

全部学科领域 搜索

首页 > 生命科学 > 期刊论文 >

Multiplicity distribution of final-state particles and different contributions of related sources in nucleus-nucleus collisions at high energies

作者	Sun Zhu; *Liu Fu-Hu	作者单位	山西大学
出处	Chinese Physics C 2008 (04) 21-22	文献类型	期刊论文
项目批准号	253486789	学科领域	D04 核物理
项目名称	兰州CSR及相关能区重离子与核乳胶相互作用研究	资助类型	面上项目
DOI	10.1029/2007JD008877	发表时间	2008
推荐引用方式	请用此识别号来引用或链接此条目: http://ir.calis.edu.cn/hdl/211010/4764		


成果所含文件  文件大小: 217 KB
下载次数: 20934 [使用许可信息请参见《国家自然科学基金委机构知识库开放获取政策》](#)

Figure 4. Layout of OA Webpage: the Third Level

4.2 Data Acquisition

There are three ways to acquire raw data and save them inside NSFC OA library.

The primary way is for the researchers from universities and institutes to input their own achievements into the database repository by themselves. In the annual, mid-term and final progress report, the NSFC project holder (PI) should fill their achievements as is required by NSFC. As is shown in Figure 5, project holder can first use the “online search” function to find their published papers from Internet and claim that achievement. This is a convenient and time saving way since author does not need to input detailed information like journal name, year, co-authors, and volume/number information. If the achievement is not formally published online, the project hold have to manually input their achievement based on certain inputting format, which means some necessary items should not be blank like journal title, authors, affiliation etc. Or else, such data might become “filthy data” since they are incomplete and not well usable later on. The third way is using “file input” function to save time. If such achievement has been input before (like inside university library) with certain format (*e.g.* txt, xls), it can be directly input just like the phone directory input function.

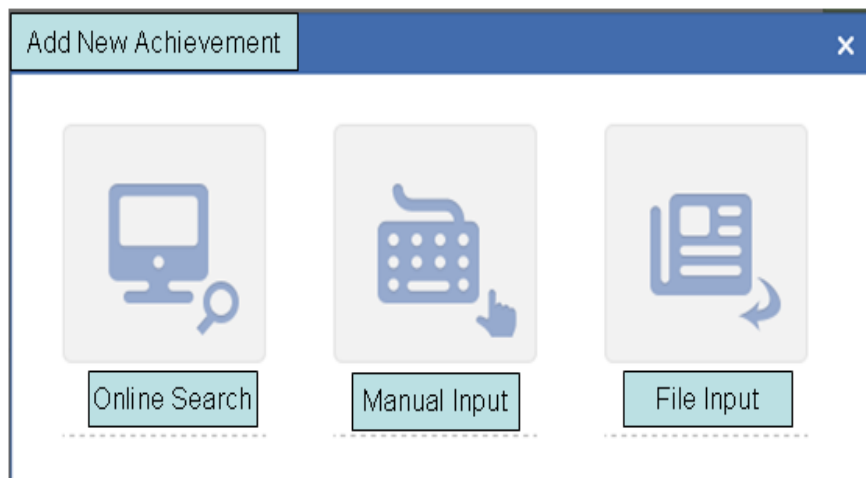


Figure 5. Three Ways to Add New Achievement

The second way to acquire raw data is to search from other famous database systems such as Microsoft Bing, Google Scholar, Scopus, Elsevier, Wiley, CNKI, Baidu etc. Such systems provides us rich and free access of many academic resources, so that we can search and periodically update the authors new achievement. If needed, we can need to buy certain database library when it is necessary.

The third way to acquire raw data is to use crawl software to obtain data from some OA repository or journal website based on their policies. It is also an important supplement way to collect raw data. Besides OA repository, from the scholars' personal homepage we can also get huge amount of valuable information like recent e-print publication etc.

It is worth noting that for the first stage of OA library construction, only journal articles are provided inside the OA repository. Later on, more achievements like patents, conference proceedings, books, dissertations and awards will be provided no matter they are projects related or not.

Also, policy support from NSFC administrative office is necessary to ensure that project holders should fill in and upload their achievement in a timely manner. System will remind them periodically to input or claim their achievements. If project relevant achievements are not reported within certain duration limit after their formal publication, some penalty rules will be applied to them such as certain minus score during final project evaluation and next time project application phase.

4.3 Key Function Modules

There are two main reasons why the middle layer key function modules are necessary for the OA library. First, there are many problems about the raw data acquired from bottom layer, which is usually called low quality "filthy data". For example, such data are incomplete, erroneous, with blank, and inaccurate. If those data are collected from various sources, there might be inconsistency about same data. Second, in order to support the upper layer efficient searching and display function, data need to be processed and managed in carefully. In the following, we will give detailed explanation of some key function modules.

Before data clearing, data analysis is need which is the precondition. By detailed and comprehensive analysis, data inconsistency, error and incompleteness can be detected. Next, data clearing can be performed either automatically or manually. Relevant data clearing rules and workflow are needed accordingly. Huge amount of clearing work can be done automatically by data clearing tools.

Data collection is used when some raw data can not be automatically corrected during input process by authors. This is because system can not find minor inputting mistakes like year, author name etc. It is very costly if direct human participation is performed. In such case, some well designed data collection tools are necessary which can detect such mistakes by using good and efficient methods such as clustering methods, rule-associated methods, and statistical methods.

Data supplement function is needed when raw data is incomplete with blank. For example, some preprint articles are lack of publishing volume, number and page information since they are not the final published. Such information needs to be supply periodically by some tools after searching from the Internet.

Before data searching, data modeling and data storage needs to be well designed. It should support both structured dataset and non-structured dataset. It should also support various access engines like HBase, HDFS, XML database and MySQL etc. with fast switch and coexistence of various storage and access methods. Data searching is a key function behind the system and for front end users. Thus, how to ensure fast, accurate and efficient searching is a technical challenge for system designers. Building of various indexing tables can help efficient raw data search.

Project holders need to periodically upload their achievements by choosing one of the three ways, as is depicted in Fig. 5 above. They will take full responsibility for achievements they uploaded, and the data quality is relatively high since they know well about their own achievements. It is worth emphasizing that for manual input case, some necessary information like author name, title, journal, year, volume/number page range etc. must be filled out to ensure data completeness.

The authors can also claim for their own achievements once they log in the OA library. System will periodically collect for the authors their achievements like joint papers or patents which they might not be aware of. Then, system will remind them via pop-up message or emails to confirm with author about their own work. In that case, authors do not need to input again such achievement which is time saving and efficient. Other co-authors can do it in the same way. So, there is only one record/item in the system while different co-authors can claim for their joint achievements.

Some policies about downloading and use of OA resources will be present during the access of OA library. For example, the users who download others' work should not misuse them to some commercial activities to make profit, nor to do crime. It should not infringe others privacy or safety from materials like e-print articles, images or audio/video documents. Records about famous stars or governmental officers are relatively sensitive information, which need to be carefully viewed and treated.

Finally, data management is the last, but not the least, important function. After a series steps of data processing, such data are viewed as clean data which can be used safely later on. How to store and organize such data in an efficient manner, how to retrieve different version articles (like e-print, revised/final version and post printed version journals) is an interesting research issue. Even though the current OA library deals with only NSFC project related journals, more achievement like patents, books, awards, dissertations etc. will be considered later on.

4.4 Other Function Modules

Other system function modules are also necessary such as system interface module, people management module and security module.

System interface module provides an interface between outside users and system data at the upper OA webpage layer. After a series steps of processing, raw data can be viewed as clean data which can be used directly by system. So, system interface module should have ability to call such data in an efficient way and provide application services to the upper layer users. The interface module should also provide other interfaces with modules like other database system, people management module etc.

People management module should have ability to add, delete, modify and find relevant people information as basic functions. It should also distinguish guest, normal user, expert and system administrator. Normal user information mainly include: Chinese and English name, gender, age, birthday, city, ID number, research area, position, academic degree, address, phone number, mail address, homepage etc. For expert or advanced user, more information is needed such as personal profile, short bio, project awarded etc. Since all users can log in the system via their own account after real name registration, it is relatively easy and convenient to management user behavior. On the one hand, system can provide users recommendation services after analyzing their interested contents. On the other hand, system can provide better upload/download service to them with security and privacy guarantee.

Security module is the last, but not the least, important function module for our OA system. Software and hardware security problems need to be thought of, such as periodical update of system. Network level security problems should also be considered like using firewall for protection from malicious attack or eavesdrop. Different users should have different levels of permission to access data after authorization. Data integrity and confidentiality should also be guaranteed via certain encryption mechanism. Finally,

based on data priority, periodical data backup should be performed both locally and remotely. For example, backup of data should be done remotely rather than locally in case certain disaster happened locally.

Some other function modules like OA advertising module, system-user interactive module, suggestion and complaint module etc. are also important, which we will not discuss here.

5. Conclusions and Future Work

We proposed the construction scheme of our NSFC open access library which is under development recently in this paper. After comprehensive survey about both domestic and abroad OA policies, OA journals and repositories, we first propose the overall architecture for OA library. Detailed design from upper layer OA webpage 3-level display, mid layer some key function modules, bottom layer raw data acquisition, and other important system function modules like interface module, people management and security modules are then explained and discussed with illustrative figures.

Future work could be extended in the following ways. First, we plan to add more project relevant achievements inside the system such as patents, conference papers, books, awards and dissertation etc. Then, NSFC project irrelevant achievements can also be included inside our OA library to promote academic exchange and sharing. Third, we can build our own expert profile system, which can use the clean data directly to better serve the expert scholars. Finally, we will improve our OA library quality in terms of accuracy, timeliness, convenience to widen its application and popularity.

References

- [1] http://en.wikipedia.org/wiki/Open_access Retrieved 1 May (2015).
- [2] "Read the Budapest Open Access Initiative". Budapest Open Access Initiative. Retrieved 1 May (2015).
- [3] K. G. Jeffery, "Open Access: An Introduction", Ercim News, (2006) January.
- [4] <http://www.nsf.gov.cn/publish/portal1/> Retrieved 1 May (2015).
- [5] <http://www.nsf.gov.cn/publish/portal1/tab157/> Retrieved 1 May (2015).
- [6] <http://www.nsf.gov.cn/publish/portal0/tab87/info44471.htm> Retrieved 1 May (2015).
- [7] <http://cnplinker.cnpeak.com> Retrieved 1 May (2015).
- [8] <http://www.doaj.org/> Retrieved 1 May (2015).
- [9] <http://www.highwire.org/lists/freeart.dtl> Retrieved 1 May (2015).
- [10] <http://www.ncbi.nlm.nih.gov/pmc/about/openftlist.html> Retrieved 1 May (2015).
- [11] <https://www.openj-gate.com/> Retrieved 1 May (2015).
- [12] <https://www.jstage.jst.go.jp/browse/-char/en> Retrieved 1 May (2015).
- [13] <http://www.dspace.org/> Retrieved 1 May (2015).
- [14] <http://repository.ust.hk/dspace/> Retrieved 1 May (2015).
- [15] <http://eprints.anu.edu.au/> Retrieved 1 May (2015).
- [16] <http://libguides.caltech.edu/CODA> Retrieved 1 May (2015).
- [17] <http://arxiv.org> Retrieved 1 May (2015).
- [18] <http://www.paper.edu.cn/> Retrieved 1 May (2015).

Author



Jin Wang, he received the B.S. and M.S. degree from Nanjing University of Posts and Telecommunications, China in 2002 and 2005, respectively. He received Ph.D. degree from Kyung Hee University Korea in 2010. Now, he is a professor in the College of Information Engineering, Yangzhou University. His research interests mainly include routing protocol and algorithm design, performance evaluation and optimization for wireless ad hoc and sensor networks. He is a member of the IEEE and ACM.