CA1678
中国における Million Book Project
－中国の大学図書館の資料電子化戦略－

1. はじめに

人類の英知をあつめた著作を後世に残し、すべての人が自由に利用できるような Universal Library を作ろうという壮大な構想がある。その実現のための第1段階として、100万冊の書籍をデジタル化し、検索できるようにして、インターネットを通じて無料提供するプロジェクト“MBP（Million Book Project）”が立ち上げられた。2007年11月に150万冊の書籍のデジタル化を終え、インターネットでの提供を開始した（E727参照）。このプロジェクトに共同で取り組んでいるのは、米国カーネギー・メロン大学と中国・インドの高等教育機関、エジプトのアレクサンドリア図書館である。

このプロジェクトは、中国では CADAL（China-America Digital Academic Library：高等学校中英文图书数字化国际合作计划）、あるいは China-US Million Book Digital Library Project と呼ばれ、大学図書館を中心とした高等教育機関のデジタル図書館の枠組みの中に位置づけられる。本稿では、中国における MBP である CADAL について紹介したい。

2. 中国での参加機関とプロジェクトの位置づけ

CADAL は、浙江大学、中国科学院を中心に、北京大学、清華大学など14機関が参加して開始された[1]。なお、CADAL 管理センターの統計では、中国人民大学、中国農業大学の2機関が増え、16機関となっている（表を参照のこと）。

中国には大学図書館を中心とした大規模な図書館コンソーシアムに CALIS（China Academic Library and Information System：中国高等教育文献保障系统：CA1443参照）がある。政府の第9次5か年計画（1995-2000）、第10次5か年計画（2001-2005）における211プロジェクト[2]の主要な公共サービスの1つで、1998年に構築が開始された、学術資源の共有のための基盤構築プロジェクトである。CALIS の機能は（1）総合目録データベースの作成・提供、（2）図書館間相互貸借・文献複写、（3）ドキュメント・デリバリー・サービス、（4）データベース構築、（5）資料の電子化、（6）インターネット・ナビゲーション・システムの構築、（7）電子的資料のミラーサイト運営、（8）資料の共同分担収集[3]等多岐にわたっており、参加機関は500を超える[4]。

第10次5か年計画の期間中の2004年には、一次資料の作成と提供を行う CADAL は情報資源の共有のためのインフラといえる CALIS とあわせて、CADLIS（China Academic Digital Library & Information System：中国高等教育数字化图书馆）と呼ばれる、CALIS の第2期プロジェクトに位置づけられた。

プロジェクトの資金については、中国教育部から7,000万元（約9億8,000万円）、米国国立科学財団（National Science Foundation：NSF）から1,000万ドル（約10億円）相当のソフトウェアとハードウェアの提供を受け、各協力機関の自己負担があわせて1,000万元（約1億4,000万円）となっている[5]。

3. プロジェクトの経過

プロジェクト開始にあたって、2000年に NSF が50万ドルを出資してスキャナーを購入し、中国とインドで大規模なスキャニングの実験を行った。その後、NSF が出資した会議に、中国からは大学、中国教育部、中国科学院の代表が、米国からは NSF、カーネギー・メロン大学からの代表が参加し、プロジェクトについての合意に達し、運営委員会の発足を決めた。2001年末には、米国の各大学が参加して、デジタル化対象資料を選択し中国やインドに移送してスキャニングをするという方針が決められた[6]。

CADAL は2002年9月に第10次5か年計画の211プロジェクトの重要プロジェクトの1つとして正式に組み入れられた。2003年9月には中国教育部から参加機関への4,000万元（約5億6,000万円）の資金提供があり、実質的なプロジェクトが開始された。2004年の8月31日までに、68,940冊の中国語の資料が電子化され、2004年の夏には、カーネギー・メロン大学の書庫から3万冊の英語の資料と政府出版物が、スキャニングのために中国に送付された[7]。

2005年11月 CADAL プロジェクトのポータルサイトが浙江大学のウェブサイトで正式公開された。この時点までにデジタル化が完了した資料は40万冊であった[8]。

2007年11月には UDL（Universal Digital Library）全体での蔵書が150万冊を超え、インターネットでの提供が開始された。そのうち、CADAL がデジタル化した資料は107万冊で、実に全体の70%以上を占める[9]。

4. デジタル化対象コンテンツ

デジタル化の対象となったのは中国語資料50万冊、英語資料50万冊のあわせて100万冊である。中国語資料については、近現代の図書30万冊以上（1949年以降に出版された学術書、図書20万冊、中華民国期の図書10万件）、プロジェクト参加機関の修士・博士論文約10万件、古籍および貴重な伝統文化資源10万件弱、マルチメディアリソースライブラリーが若干という構成である。

⑴ 近現代の中国書

1919年以降に出版された、主要な学術著作と数学、科学研究に重要な参考図書を対象とし、原則として小説などの文学作品を含まない。北京、上海で出版された出版物を中心に、大学出版社等の学術書、省レベルの人民出版社、科学技術出版社の図書が対象となっている。小中学生向けの児童書、小説、散文、ルポルタージュ、詩歌、演劇などの文学・芸術作品、人物の伝記、150ページ以下の図書、2000年以降に出版された図書および翻訳書は含まれていない。

1912～1949年の中華民国期に出版された資料は10万タイトルを超え、著作権の保護期間を過ぎている資料も多く、比較的まとまったコレクションとなっている。中華民国期は中国の歴史上、中華文明と西洋文明との衝突により、多くの学術的・歴史的な資料が生み出された。その一方で、出版技術は伝統的な技術からの過渡期にあり、伝統的な製紙・製本技術で作成された図書と西洋の技術による洋装書が混在する。とりわけ当時の洋紙は酸性度が高く脆弱であることから長期保存に適さず、デジタル化の必要性の高い資料群といえる[10]。

⑵ 学位論文

CALISの学位論文データベースの構築と相互に連携し、参加機関の1980年代以降の著作権が明確な修士論文、博士論文計約10万件のデジタル化を進めた。

⑶ 古籍

CADALでは、1911年以前に出版された資料を古籍と定義し、中国に現存する古籍の多くを含む『四庫全書』、『続修四庫全書』[11]から選択してデジタル化を行った。この他、北京大学からは宋代、元代の貴重書、家譜、絵画、碑文の拓本、清華大学からは古代の科学技術史とその関連文書、南京大学からは太平天国の関連史料、浙江大学からは敦煌文書や茶文化に関する資料、帛書、帛画など、それぞれの協力機関の所蔵資料がデジタル化の対象となった。

⑷ マルチメディアリソース

書籍以外では、「中華文化の保護」をテーマとし、書画、音楽などのデジタル化を進めた。

⑸ 英語資料

約50万冊にのぼる英語資料は、米国側から提供された。1）電子図書館連合（Digital Library Federation：DLF）に所属している大学図書館の著作権が明確な所蔵資料、2）政府出版物や保護期間が過ぎた資料など著作権のない資料、3）学位論文、技術報告書、会議録などのボーンデジタルの資料で、DLF所属の25の大学の学位論文の全文、出版社と交渉中の10万冊の電子書籍等が含まれる。

5. 著作権処理

CADALでは著作権が明確なもの、あるいは経済的な利益追求の度合いが比較的低い学術書をデジタル化の対象とし、文学作品や新しく出版された図書を原則的には対象から外すことで、著作権の許諾を得るコストを引き下げている。著作権処理は、1）著作権者に有利な条件を提示して直接交渉する、2）デジタル化の対象となる資料のリストを公開する専門のポータルを作成し、著作権者にデジタル化の権利の提供を求めるなど、メディアを通じて呼びかけを行う、3）中国著作権保護センターと協議し、デジタル化の対象となる資料の著作権管理の一部を同センターに委託するという3種類の方法で行っている。

CADALの管理センターのサイトでは、出版社向けのメッセージが公開され、絶版だが著作権は存続している出版物について、CADALでのデジタル化と提供についての許諾を求めている。出版社に対しては、CADALがデジタル化したデータを出版社に提供することで、オンデマンド出版の販売も可能になること、他機関がCADALのデータを商用利用する際には、著作権者に一定の費用を支払うよう求める権利があること、減税の措置があること、などのメリットを挙げている。また、CADALでの提供はいつでも取り消すことができる柔軟な仕組みを提示している。ただし、CADALプロジェクトのウェブサイトから該当のデータを削除するには、スキャニングと保存に費やした費用（1冊につき200元（2,800円））を支払う必要があるとしている[12]。

6. デジタル化のプロセス

CADALでは、14の協力機関にそれぞれデジタル資源センターが設置され、資料を提供する機関がデジタル化を行い、データを提供している。技術的な支援については、浙江大学と中国科学院が中心となっている。米国から送付される英文資料については、深圳の保税区内に毎月2万冊のペースでデジタル化が可能な施設を設立して作業が行われた。

2008年6月28日現在の統計によれば、デジタル化の冊数は浙江大学が約66万と飛び抜けて多く、北京大学、復旦大学が約10万冊、南京大学が約7万冊、中国人民大学、四川大学、清華大学、武漢大学が5万冊前後と続く。コンテンツ毎の内訳では、学位論文については、中国科学院文献中心、中華民国期の雑誌や図書については復旦大学、古籍については北京大学、現代図書については浙江大学のデジタル化冊数が最も多くなっている（表を参照）。

表　CADALの機関別デジタル化冊数
（2008年6月28日現在。単位：冊）

	学位論文	中華民国期図書・雑誌	古籍	現代図書	英文図書	その他	総計
北京大学	0	1, 119	106, 372	0	0	626	108, 117
精華大学	10, 286	23, 118	16, 945	0	0	7	50, 356
浙江大学	26, 610	11, 135	55, 992	418, 461	147, 913	2, 262	662, 373
南京大学	1, 968	52, 860	8, 938	9, 316	0	0	73, 082
復旦大学	0	90, 334	0	14, 160	0	0	104, 494
中国科学院文献中心	38, 830	1, 000	0	2, 198	0	1, 600	43, 628
上海交通大学	10, 757	1, 287	0	949	2, 463	0	15, 456
西安交通大学	10, 944	5, 376	4, 004	5, 155	14, 145	0	39, 624
武漢大学	16, 945	17, 799	0	1, 182	14, 274		50, 200
華中科技大学	18, 522	1, 730	0	1, 400	9, 317	0	30, 969
中山大学	0	0	0	0	15, 467	0	15, 467
吉林大学	21, 848	24, 308	307	0	0	0	46, 463
四川大学	10, 649	12, 585	3, 927	0	30, 100	0	57, 261
北京師範大学	8, 205	19, 176	10, 250	4, 668	0	0	42, 299
中国人民大学	9, 212	49, 974	0	581	0	0	59, 767
中国農業大学	4, 695	3, 508	0	2, 182	3, 535	0	13, 920
総計	189, 471	315, 309	206, 735	460, 252	237, 214	4, 495	1, 413, 476

※高等学校中英文図书数字化国际合作计划项目管理中心．“数字化进展”．
http://www.cadal.cn/cnc/cn/xmdt/szhjz.htm,
（参照 2008-09-06）を参考に作成。

デジタル化の解像度は、通常は600dpiの2値、グレースケールのページがある場合は600dpiで256階調、カラーページについては600dpiのトゥルーカラーが採用された[13]。提供にあたってはDjVu形式の画像ファイルに変換され、全文検索のためにOCRで読み取ったテキストデータを格納している。

なお、ウェブサイト上でもデータの誤りが見つかった場合に閲覧者からフィードバックをうける仕組みを設け、データの品質向上に努めている。

7. メタデータ付与

CADALで提供される中国語の図書、古籍、学位論文のメタデータについては、浙江大学が北京大学、上海大学他と協力して『CADAL数字化文本元数据规范草案：Edocument Metadata Version2.0』[14]を定義している。OEBPS（Open eBook Publication Structure Specification）、Dublin Core、MARC XML Schemaなどを参照して作成されているということである。また、この定義に加えて古籍記述細則、民国図書記述細則、普通図書記述細則、雑誌メタデータ記述細則などの記述細則を作成し、適用している。

また、書誌事項についてはOCLCやCNMARC（中国国家図書館が製作している国家標準のMARC）の既存の書誌データを変換して流用している。

8. データ保存とアクセス

CADALには1冊あたり約50～60MBの画像とテキストファイルが収録され、総データ量は3億ページ、6,000億字にのぼるとされている[15]。大量のデータを扱うため、複数のミラーサイトを設置し、アクセス速度の向上、セキュリティと長期保存を保証している。UDL全体では、米国、中国の南北2か所のセンターの他、インドでもミラーサイトが提供されている。

CADALプロジェクトでは、E727や、またレディ（Raj Reddy）の論考[16]が指摘しているようにプロジェクトの開始当初から事業の継続の可能性についての議論がなされているが、現在のところ、今後の具体的な計画については不明である。しかし、100万冊を超える規模でのデジタル化を特定の国立図書館など1館が行うのではなく、複数の図書館で分担して実施したこと、また資料の提供とデジタル化を別々の国で担当するなど、国際的にも多くの機関が共同して行ったことは、今後大規模なデジタル化を行う際のモデルになりうるという意味で一定の成果があったといえる。事業の継続展開が望まれる。

（関西館アジア情報課：篠田麻美（しのだあさみ））

⑴［高等学校中英文図书数字化国际合作计划项目管理中心］.“百万册书数字图书馆项目在中国的背景情况”.
http://www.cadal.net/cnc/cn/cncadal.htm,（参照 2008-09-06）.
⑵ 21世紀に向けて100あまりの大学を重点的に育成、発展させるという国家プロジェクト。
⑶ 呑海沙織．中国における学術図書館コンソーシアムと電子図書館プロジェクト－CALIS，CADALからCADLISへ．日本農学図書館協議会誌．2005,（136）, p. 9-13.
⑷ 中国高等教育文献保障系统管理中心．“成员馆”．中国高等教育文献保障系统．
http://www.calis.edu.cn/calisnew/calis_index.asp?fid=14&class=5,（参照 2008-10-06）.
⑸［高等学校中英文图书数字化国际合作计划项目管理中心］.“资金来源”.
http://www.cadal.net/cnc/cn/zjly/zjly.htm,（参照 2008-09-06）.
⑹ Reddy, Raj et al. The Million Book Digital Library Project. 2001-12-01.
http://www.rr.cs.cmu.edu/mbdl.doc,（accessed 2008-10-06）.
⑺ Zhao, Jihai.“Technical Issues on the China-US Million Book Digital Library Project”. Digital Libraries: International Collaboration and Cross-Fertilization: 7th International Conference on Asian Digital Libraries, ICADL 2004. Shanghai, 2004-12-13/17. Springer, 2004, p. 220-226.
⑻ 单冷ほか．数字图书馆门户网站在浙大开通．光明日报．2005-11-03, p. 2. 入手先，中国重要报纸全文数据库，
http://cnki.toho-shoten.co.jp/kns50/Navigator.aspx?ID=3,（参照 2008-10-06）.
⑼ 周炜ほか．全球数字图书馆扫描图书突破150万册 中国贡献七成．光明日报．2007-12-04, p. 2. 入手先，中国重要报纸全文数据库，
http://cnki.toho-shoten.co.jp/kns50/Navigator.aspx?ID=3,（参照 2008-10-06）.
⑽ Zhan, Meng et al.“CADAL and the Literature of Republic of China”. ICUDL 2006. Alexandria, Egypt, 2006-11-17/19, Universal Digital Library.
http://www.ulib.org/conference/2006/21.pdf,（accessed 2008-10-06）.
⑾『四庫全書』は清の乾隆帝（1711-1799）の勅命によって編纂された中国最大の叢書である。古今の重要な書物約7万9千巻余を、経・史・子・集の四部に分類し、収録している。『続修四庫全書』（上海古籍出版社，1995年）は『四庫全書』未収の5,000余種1,800冊を収録。
⑿［高等学校中英文图书数字化国际合作计划项目管理中心］.“CADAL公开信”.
http://www.cadal.net/cnc/cn/bqgg/letter.htm,（参照 2008-09-06）.
⒀ CADAL项目管理中心．CADAL数字化文本加工规范草案：

Edocument Digitization . Version2.0, 2004.
⑭ CADAL项目管理中心. CADAL数字化文本元数据规范草案: Edocument Digitization . Version2.0, 2004. http://www.cadal.net/cnc/cn/jsgf/CADAL_metadata_2004.pdf, （参照 2008-09-06）.
⑮ Zhao, Jihai. “Technical Issues on the China-US Million Book Digital Library Project”. Digital Libraries: International Collaboration and Cross-Fertilization: 7th International Conference on Asian Digital Libraries, ICADL 2004. Shanghai, 2004-12-13/17. Springer, 2004, p. 220-226.
⑯ Reddy, Raj et al. The Million Book Digital Library Project. 2001-12-01. http://www.rr.cs.cmu.edu/mbdl.doc, (accessed 2008-10-06).

Ref.
高等学校中英文图书数字化国际合作计划 . http://www.cadal.zju.edu.cn/,（参照 2008-10-06）.
Carnegie Mellon University. “The Universal Digital Library”. http://www.ulib.org/,（accessed 2008-10-06）.
陈海英ほか . 中美百万册数字图书馆项目综述 . 大学图书馆学报 . 2005, (1), p. 3-6,13. 入手先 , 中国期刊全文数据库 , http://cnki.toho-shoten.co.jp/kns50/Navigator.aspx?ID=1,（参照 2008-10-06）.
Carnegie Mellon University Libraries. “Frequently asked questions about the Million Book Project”. http://www.library.cmu.edu/Libraries/MBP_FAQ.html,（accessed 2008-10-06）.