国際連携漢籍資料庫の夢

—漢籍のデジタル化について—

丘　山　　新

（東京大学東洋文化研究所教授）

こんにちは、丘山です。私は中国、インドの仏教史を研究していますが、同じ中国研究と言っても思想的な研究でいわゆる漢籍とは違う範疇にあり、漢籍の専門家ではありません。また、別にネットワークやコンピュータなどに強いわけでも全然ありません。どちらかというと両方ともに素人で今日のテーマからすると少々場違いな気もするのですが、研究所の中でいつの間にか漢籍データベースに関わることとなり、そんな次第で話をさせていただくことになったわけです。

私はこちらの研究所に来た以上、自分の研究だけではない仕事を何か一つしたいと思い、4～5 年国内のいろいろな状況を見渡しておりました。すると、漢籍の目録データベースの構築に関して、どういう項目を取るかとか、今 Unicode で入力するようになった漢字のコードについても、当時は台湾の Big5、日本の JIS コード、中国大陸の GB コードなど、いろいろな可能性があって、どれがいいのかということを、日本のいろいろな大きな会議で 10 年ぐらい議論ばかりやっていて、なかなか実際の漢籍目録のデータベースができそうにない状況なのが分かってきました。幸か不幸か、私はそういう世界と全然関係なく孤立して研究していましたから、それらの議論とは関わらず、とりあえず試作版を作ってみようと思いました。データベースの考え方とか、コンピュータのいろいろな技術はどんどん進んでいきますから、どれがいいかなどというのは幾ら議論しても、例え 10 年かかっても、これが最終的で最善な手段だとはならないと割り切って、そのときの最新技術と一番いいであろうという方針でやればいいのだと考えたのです。

本研究所は、田中所長が挨拶で申したように、教員が主体になって、それを事務方が応援するというかたちで様々な事業を進めてきました。この漢籍目録データベースに関して言いますと、私はいろいろな技術に関しても漢籍に関しても素人です。ですから、東洋学研究情報センターと図書室などの事務方の人たち、それからアルバイトの人たちに、実際の仕事を含めた多くを頼ることになり、私はアイディアを出したり資金集めなどの小間使いのようなことをしたりしただけで、本当にそんな状況でやってきたわけです。

東洋文化研究所には東アジアから西アジアまでのいろいろな研究者がいるのですが、その関係で資料データベースは随分いろいろできてきています。今年新しく予算がついて、本日の司会の西アジア研究部門の鎌田先生が関係されているアラビア語のダイバー・コレクションのデータベースを来年４月から公開します。それから、総合図書館の仏教の古い写本の資料のデジタル化を、南アジア研究部門の永ノ尾先生が作成されることになっています。この研究所では領域や対象地域、それから研究の分野が本当に様々で、国際政治専門の田中所長などは「世界と日本」という非常に高い評価を受けているデータベースを、強力なスタッフと協力して作っています。ばらばらに見えるかも知れませんが、それは研究領域が大変広いということでもあるのです。資料のデータベース化は義務ではありませんが、関心を持っている教員が携わって多様なアジアに関するデータベースができてきました。

データベース化は最近ではどこの研究機関でも取り組んでいますが、私はとにかく人がやっていないことをやりたいと考えてやってきました。漢籍目録のデータベース化はほぼ完成してきているので、２年ほど前から研究所に所蔵されている貴重漢籍の全文画像データベースを構築し始め、既に公開もしています。これに関しては、国際的に見ても、せっかく作ったものを非公開にしている機関がほとんどで、それをとても残念だと思っています。私どものところでも全面公開にはなっていませんが、出来れば世界中の機関が持っている貴重資料のデジタル資料庫を作って、資料に関しては世界のどこの研究者も平等に扱えるようにしたいと思っています。東洋文化研究所が、アジア資料のデータベース化のために、世界に先駆けて試行的なものを開発していくのが、私の夢なのです。

さて、今日の講演会のテーマはどのように保全していくかということが主題ですが、私は研究者の立場から、どういうものが欲しくてどのように使えるようにしていくか、ということをお話させていただきます。

東洋文化研究所漢籍目録データベース（図 1）は、先ほどお話したように「えい、やってしまおう」と 1997 年に企画しました。企画したといってもまず個人で始めたわけですが、

研究所の冊子体の漢籍目録これ 1 冊抱えて台湾に行き、知り合いを通して評判のいい業者を紹介してもらってどのくらいの価格でやってくれるかと打診したところ、日本の業者の提示価格のなんと 1/5 ぐらいというので飛びつきました。ただ、当時は台湾でもそういう目録データベースが出来ていなかったため、台湾の業者にも「本気か?」などと言われたものですが、翌年科学研究費を取ることができて実現しました。このような作業を外注でするか、自分たちでするかというのは問題となるところで、後でもう少し触れるつもりでいますけれども、基本的な部分は外注して効率よくやってしまったほうがいいと私は考えています。

図 1. 東京大学東洋文化研究所漢籍目録データベース
(http://www3.ioc.u-tokyo.ac.jp/kandb.html)

1999 年、試験段階で最初 1 万件ぐらい公開しましたが、そのときは Big5 という台湾の漢字コードを使っていました。当時は Unicode をまだ使えなかったのです（今年からこの目録ＤＢも Unicode 版で公開しています)。図書室の人たちからは、Big5 をインストールしていない環境では使えないだろう、中国語読みできない人はどうするのかなどと後々しばらく言われました。そういうときに私はいつも開き直って、「私は国内向けなどとは考えていません」と申しました。最初から国際的なものを作るのが夢でしたし、その当時 Big5 というのはいちばん漢字の数も多くて目録の表記にはちょうど向くと考えたのです。しかも世界の中国研究者は、当時は Big5 をどこでも使っていましたし。

とにかくデータベースを作るときはその時点で一番適切だと思える方法で、なおかつ将来

的にこういうふうになっていくだろうと考えて、将来そこをうまく変換できるように事前にいろいろな仕組みを作っておけば、幾らでも対応できるのです。基本的に私には「データベースは生きているのだ」という思いがあります。生きているから、成長もしていくし、いろいろ世話をしていかないと途中で死んでしまうこともあると考えています。

これは少し宣伝になりますが、この目録データベースを台湾に持っていったら、台湾国家図書館や中央研究院など、いろいろなところの人が、ああしたらいい、こうしたらいいと、すごくいろいろなアドバイスをしてくれました。2000 年に台湾も一挙に台湾地区の連合目録を作っています。それをご覧になると分かりますが、私たちのところと非常に見かけが似ています。彼らは私たちのデータベースを試験台にして 1 年後に自分たちのものを作っていったのです。私はそういうのをずるいとは全然思っていなくて、大いに使ってくださいと考えています。そのようなきっかけになれたら、それでいいことだし、光栄なことでもあると思います。その後、この台湾地区の漢籍の連合目録側からはずっと一緒にやろうと言われていまして、やっと今年度内には国際的な連携を実現させられると思っています

さて、日本の大学機関では漢籍関係に強いところでは、京都大学の人文科学研究所があります。人文科学研究所でも本当はこういうことは国内で一番先にやりたかったと思われるのですが、こちらが先行しましたので急いでやらなければいけないということで、協力を求められました。そういうことならと東洋文化研究所ではデータを全面的に提供して、2001 年から、国立情報学研究所と人文研と東文研とが幹事機関となり、全国版の漢籍目録のデータベースが構築され、今のところ収録データは 60 万件ぐらいという話です（東文研分は大体 10 万件近くです)。

さて、漢籍目録データベース立ち上げから 5〜6 年経って、全国版もできることだしもうこれで引退しようかと考えたのですが、まだ漢籍の全文を画像で公開しているところがありませんでしたから、原資料の補修や保存とも関連させて漢籍のデジタル化事業を進めることにしました。貴重な漢籍に関しては、原資料は保存に努め、代替資料に関しては、今まではマイクロ撮影を行ったりそのデータから複本を作成したりしていました。漢籍でいうと、叢書類に関しては影印本がけっこうたくさん出来ているのですが、そういう紙資料とは別にデジタル資料を作ろうと考えたのです。デジタルデータがどの程度の期間もつのかよく分かりませんでしたが、所蔵機関に足を運ばなくても世界中どこでも見られるという意味では、デジタル化は研究者には非常に便利だろうと考えたわけです。

このときも先ほどお話しました皆さんの尽力があり、無事予算を獲得できました。積極的

に新しいアイディアを出して、それが有用であれば大体お金はついてくるのではないかと私は甘く考えています。ただ、科学研究費補助金などを申請するとき、私は教員だけではなくて実際に資料に携わる図書室関係の人たちなどもメンバーに入ってもらいたいと主張したのですが、今のところ文科省でも日本学術振興会でも、そういうスタイルをまだ認めていないのです。これは大きな問題だと思います。やはり実質的に関わってくださっている方々の業績としても、きちんと認めてもらえるようなシステムを作るべきだと思います。私はもちろん台湾に行って向こうの人たちにも協力をお願いしたりしますが、実作業面では今日もここにいるアルバイトのスタッフにほぼ全面的に任せているのです。

さて、本研究所の漢籍善本全文影像資料庫（図 2）は、資料全文をデジタル化したものです。いわゆる電子図書館を構築している研究機関の中には、ごく少数のきれいなカラー画像を公開しているところもあると思いますが、私たちのところの基本的な考えは違っていて、研究のために、特に私個人の願いとしては全部公開して見せてしまえ、ということです。ですから、予算が許す限りどんどん出来るだけ多くの資料をデジタル化しなければいけないと思っています。カラー画像ではなく白黒画像で作製しているのも、そういう理由です。で、とりあえずは、特別貴重書、と私どもが所内で呼んでいるものからデータベース化しています。

図 2. 東京大学東洋文化研究所漢籍善本全文影像資料庫
(http://shanben.ioc.u-tokyo.ac.jp/)

それから、今このようにしたいなと思っているのは、国内外の善本影像をデジタル化しているところは全部お互い一斉に公開して、あるいは先ほど申し上げましたように、一つの大きな善本のデータベースを仮想空間上に作って、それをみんなで、資料に関しては平等に用意ドンで研究できるようにするということです。それが結果的には、貴重な漢籍などを保存するのにも役に立つであろうと思っています。

ところで、データベースというのは極端にいえば外注に出せばできると批判されることもありますが、そういう批判はあたりません。勿論いくら外注してもそのままで公開できるわけではないのです。力の配分でいえば、外注の力が４割、戻ってきたものをまた工夫していろいろするのに６割ぐらいの力は要るので､簡単に丸投げをやっているわけではありません。こちらでも、いろいろなアルバイトの人たちが本当によくやってくれているのです。特に、データベースの正確度からいうと、構築作業の過程で校正は何より大切なのです。

さて、漢籍の資料庫と知識庫、というようなことを私は考えています。資料庫というのはとにかく使える資料をデータ化する。資料蓄積庫です。それをベースにして、主として研究者がいろいろなアイディアを出し、工夫をして付加価値を加えながら、いろいろな新しい研究データベース、知識庫を作っていくのだと思います。この知識庫構築が、今後のデータベースの基本思想になると、私は考えています。

善本全文影像資料庫の今後の課題ということでは、すでに収録した 566 点にプラスして今年度収録タイトルを 600 点ぐらいにできると思っています。ただし、諸般の事情により、全面公開しているのは 450 点ほどで、残りのものは最初の 10 ページしか見られません。東洋文化研究所に来ていただければ全部見られますし、あと、学内の総合図書館などでは全部見られるようにしてありますが、できれば世界中に全面公開したいと私は個人的には思っております。いずれは必ずそうします。

さらに、全文テキストデータとの連携の話もあります。画像の場合、テキストデータではないので、一字一字を検索したいと思ってもできないのです。最近北京で開発されたものがあって、影像テキストも、かなりの精度のよさでテキストデータに簡単に変換できるようになっているので、そんなことも考えていきたいと思っています。それから、画像の国際共通化に関しては目録データベースほど細かい話は不要と私は思っているのですが、やはり緩やかな規定というか、標準化の話し合いも必要でしょう。これは本格的に来年あたり、とりあえず国内と台湾あたりを相手に実現できたらいいなと思っています。

私自身のこれまでの経験からの教訓としては、実現は無理だと思われても、夢を唱え続け

ていれば、大体3年ぐらいすればそれを実現できてくるという印象があります。今日の標題に「国際連携漢籍資料庫の夢」と書いておきましたけれども、3 年ぐらいすれば夢ではなくなって、「夢」が取れるのではないかと思っています。

理念的にはこんなことでやってきているわけですが、この場をお借りして改めて申しますと、おかしなことなのですけれども、本当に私はあまり何も知らないのです。事務方やアルバイトの方々に支えられてというか、頼りながらやってきているというのが実情です。そういう皆さんに、私はこの場を借りて、改めて感謝したいと思います。

（ここで、図書室職員による漢籍目録データベースと漢籍善本全文影像資料庫の操作説明）

（Q）　今日はありがとうございます。デジタル化、画像の取り込みに関して、実務的なところでお伺いしたいのですが、作業を進めていって、「この解像度では結局だめだった」という問題が起きたら、もう一回やり直しになってしまうと思います。実際、スキャナーを使うのか、写真で撮っているのか、実際の解像度をどう処理しているのかというところを、参考までにお聞かせいただければと思うのですが。

（丘山）　基本的には、カラーページは所内で写真技術者に撮ってもらっています。それから、白黒のものは基本的には外注です。一挙にデジタル化してもいいのですが、データがどのぐらいもつか、それから、マイクロフィルムとどちらの保存期間が長いのかまだ分からないので、まずマイクロをとって、なおかつそれをデジタル化しています。それから、なぜ全文全ページカラーにしないかというと、とにかく大量に公開したいという目標があるからです。カラー撮影ではお金がかかるということで白黒にしています。

解像度に関しては、もともとのものは高い解像度で撮影していますが、公開する段階でいろいろ問題があります。利用する場合、画像のダウンロードに時間がかかりすぎないような範囲で解像度を出来るだけ高くします。さらに PDF ファイルでも表示できるようになっているので、細かい部分でも表示できないことはほとんどないと思います。ただし、あまりこういうものを使って商売をしようという人はいないと思いますが、そのまますべてのデータをダウンロードして、製本出版してもきれいな本はできないぐらいの精度には落としてあります。また、解像度については外注する場合の単価価格の問題もあります。数値に関しては、今日の資料に細かい数字が入れてあるので、ご覧ください。

ただ、いろいろな会議に呼ばれて話しをすると、皆さん解像度を気になさるみたいですが、基本的にはそれぞれ予算の許す範囲でとか、それから、通信速度などはどんどん良くなってくるだろうから解像度は高くてもいいだろうとか、割り切って自分なりの基準を作ってしまえばよいのではないかと思います。画像のデータベースを多数の機関で連携させていく場合でも、解像度は統一しなくても支障は出ません。それともう一つ、私が気楽に考えているのは、駄目だったらまたやればいいだろう、データベースはそんなものだろうということです。データベースは作った後でもある程度変えていけますし、原版は解像度をよくしておいたほうがいいのでしょうけれども、そこで迷ってなかなか進まないよりは、どんどんやってしまったほうがいいのではないかと思っています。

（Q）　まず、先ほどから善本とか貴重書という言葉が何度も出てきたわけですが、その基準はどのように定めていらっしゃるか、お伺いしたいです。

それから、現在、善本全文影像資料庫に入るためには、まず東文研の目録データベースを経由することになるわけですが、連合目録である全国漢籍データベースの検索を行って、その検索結果からこの善本の資料庫に飛ぶことはお考えではないのでしょうか。

もう一つ、台湾、中国、日本という形で、国際的な漢籍のデータベースを構築していくということ、これは夢ということなのでまだ将来的なお話だと思うのですが、例えば日本側で貴重な物もどんどん公開してしまう。台湾、中国は、それに見合うだけのものを出してくれるという見込みのようなものはあるのでしょうか。現状で台湾の中央研究院などは非常に多くの画像を持っていても、それは海外からは有料だとか、アクセスのハードルを非常に高くしているという状況があります。それに対してはどのようにお考えでしょうか。

以上 3 点、お答えをお願いいたします。

（丘山）　私どもの今の基準では、善本　特別貴重書としたもので、1735 年以前までのものが特別貴重書です。それを先ずデジタル化しているわけです。それ以降、1911 年以前の主に線装本も貴重書としていますが、ここまで含めると点数がとても多くなってしまうのですべてをデジタル化するのは大変ですが、しかし私はそれは夢だとは思っていません。とにかく、線装本は再彫されることのない、いわば文化財的なものなのですから、利用にはデジタル化されたものを利用し、原資料は保存に努める、というのが基本方針です。それから、貴重書以外でも、傷みが激しい物や利用の多い物を、今後は優先してデジタル化していこうと考え

ています。

第 2 点に関してですが、目録からではなくて、東文研や研究所に附置された東洋学研究情報センター、あるいは図書室のホームページから、直接、善本全文影像資料庫に入れるようにもなっています。全国版から入れるようにできないのかというご質問ですが、私どもとしては、それはやれればいい、できれば早急に京都と相談して実現したいと思っています。私どもとしては、どんどん公開したいという立場です。また、利用する立場から公開できるものはどんどん増やしていきたいと考えています。

第 3 点の公開の問題です。実は私は、これは国際的にも各研究機関が今抱えているいちばん大きな問題だと思っています。中国、台湾、日本にかかわらず、やはり私としては、どこの機関も公開していただきたい。そういうことを話し合う国際会議も開かれています。ただ、私の個人的な考えでは、どこもやっていなくても東洋文化研究所では公開してしまえ、そうすれば相手に対して強く言えるだろうとも思います。なぜ資料を囲みたいのか、公開したくないという考え方を、私は理解できません。ただ、現実には、囲い込みをしたいという考えもあるし、どんどん積極的に公開したいという意見もある。その葛藤は、本研究所にも正直ないわけではありません。

それから、今のところ現実には東洋文化研究所でも一部は非公開になっています。これは、画像データベースを構築していながら、全面非公開になっている研究機関と交渉するために、そうしているのです。全面公開が無理であれば、取りあえずは互いに公開しあう、そういう機関が増えてきた段階で、一般にも全面公開していこう、そういう戦略から、一部非公開にしています。

それから、有料化に関しては、いずれ多分あまり意味がなくなると思います。お金を取ろうというのはごく短期的な今の考えで、私の予測では、多分 10 年かからずに、そういうものも無料化すると思います。現に、有料化しているデータベースにはアクセス回数が減少し、そういう意味では評価も下がっていくのが現実なのですから。

東洋文化研究所シンポジウム
アジア古籍保全のために

平成１７年１２月１６日

国際連携漢籍資料庫の夢
ー漢籍のデジタル化についてー

東京大学東洋文化研究所所蔵
漢籍目録データベースと漢籍善本全文影像資料庫

東洋文化研究所
図書室と東洋学研究情報センター
丘山新

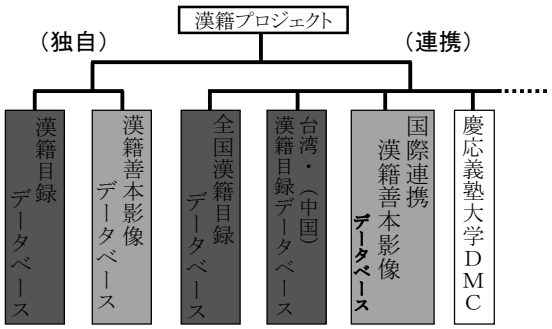
東洋文化研究所
図書室と東洋学研究情報センターの漢籍事業

- 東文研所蔵漢籍目録データベース
  - 1997年　企画
  - 1999年　Big5版　試験公開
  - 2000年　台湾地区連合目録データベースの構築
  - 2001年　全国漢籍データベースの構築
    情報研・人文研・東文研
  - 2005年　Unicode版を公開
- 東文研漢籍善本影像資料庫
  - 2005年　試験公開

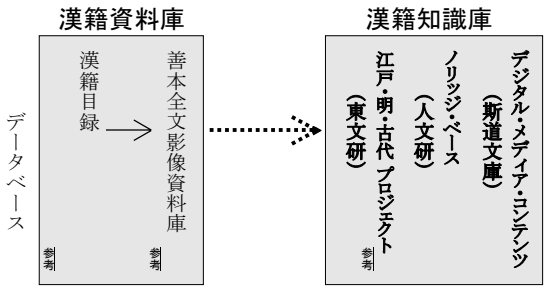
漢籍善本全文影像資料庫の構築

- 2003（H15）～
  センタープロジェクト
  「貴重漢籍の補修とデジタル化」開始
  法人化→所蔵資料は原則公開
  →代替資料作成による原本の保護
- 2004（H16）～
  基盤A「アジア古籍電子図書館の研究」
  DB科研「アジア多言語デジタルライブラリ」

東洋文化研究所
図書室と東洋学研究情報センターの漢籍事業



漢籍総合データベース



今後の課題

- 収録タイトルを増やす (現在422／566点)
- 公開タイトルを増やす (現在398／566点)
- 全文テキストデータとの連携
- 国際連携のための「ゆるやかな」標準化の制定

→ネットワーク上に「国際漢籍善本影像
連携資料庫」を構築

漢籍目録データベースと善本全文書影資料庫　参考資料

東洋文化研究所　漢籍の性質：

１．東京大学東洋文化研究所には、旧東方文化学院蔵書、大木文庫、倉石文庫などを中心に、総計約１０万点の漢籍が所蔵されています。これらのうちには孤本・善本も数多く含まれ、質量ともに世界で有数の漢籍コレクションの一つです。

２．文化財としての漢籍善本の保存をおこないつつ、多くの研究者の研究に資するため、資料をネットワーク上で試験的に公開しています。

保存作業：

貴重漢籍複製化プロジェクトとして、特別貴重書（東文研内専門家と東文研図書室の選定基準に従った三百数十種）を優先的に、ＭＦと複製本とデジタルデータを作成し続けている。現在は、特別貴重書の撮影を一部を除いて終えており、続行して、特別貴重書以外の優先順位の高い漢籍を選定し、その撮影に入っている。

（特別貴重書の中で一部未撮影のものとは、修復が必要なもの、一タイトルが大量で、使用頻度を考えると他の漢籍を優先してするべきだと判断されたものである。）

保存作業の流れ

（１）ＭＦ撮影

35mm　ネガフィルム、業者５０～１００年保障

（２）複製本作成

撮影したマイクロフィルムを元に作成。

（３）デジタルデータ作成

【保存用画像】400dpi、A3 版認識、TIFF データ

【ウェブ公開中画像】

本文（モノクロ）

小　Jpeg 画像　解像度：120dpi　幅：約　550pixel(成行)　　高さ：　500pixel

大　Jpeg 画像　解像度：120dpi　幅：約 1300pixel 強(成行)　高さ：　1200pixel

カラー巻首項

小　Jpeg 画像　解像度：　72dpi　幅：約　200pixel(成行)　　高さ：　500pixel

大　Jpeg 画像　解像度：　72dpi　幅：約　650pixel 強(成行)　高さ：　1200pixel

（4）ウェブ上公開

東洋文化研究所漢籍善本全文影像資料庫

http://shanben.ioc.u tokyo.ac.jp/

・ 現在、５３７タイトル（一部作業中）公開中

・公開範囲

A 群：全公開

B 群：制限公開

全公開　：東洋文化研究所内・東大総合図書館・東大駒場図書館・本郷文学部内

制限公開：上記以外の場所から閲覧する際、はじめの１０ページだけが表示可能

（A 群、B 群の区別は一覧表に現れておりません）

目録データベース：

東洋文化研究所所蔵漢籍中、約８万点、入力を終えており、漢籍の検索可能となっております。このうち特に、特別貴重書と指定しました、約５３７タイトル（年度末までにはさらに件数が、増えます。）は、善本全文影像資料庫へのリンクを貼り、インターネット上で漢籍書影の閲覧を可能にしました。これによって古籍原本保護と資料利用の範囲を広げることを実現しようとしています。

また当研究所では、原本保護のために、影印本や排印本がある資料は、できるだけ影印本や排印本の閲覧をお願いしています。そのためにも近年刊行収蔵された《續修四庫全書》《四庫全書存目叢書》《四庫未収書輯刊》等、影印本の大型叢書の入力作業の計画を進めております。現在は、《續修四庫全書》の来年度からの公開を目指し作業中です。

東洋文化研究所所蔵漢籍目録データベース

http://www3.ioc.u tokyo.ac.jp/kandb.html

・約８万点、入力完了

・Unicode フォント使用のため、漢字は基本的に、どのコードで入力しても検索可能

・動作環境

WindowsXP	Macintosh OS X
・Netscape7.1 ・InternetExplorer6.0 ・Mozilla1.7.5（FireFox 非対応）	・InternetExplorer for Mac 5.2 ・Safari 1.2
※ IE 以外では表示のバランスが崩れることがあります。	※ 分類検索は正常に動作しないことがあります。

皆様のご協力をお願いいたします。

平成 17 年 12 月 15 日