資料2－5

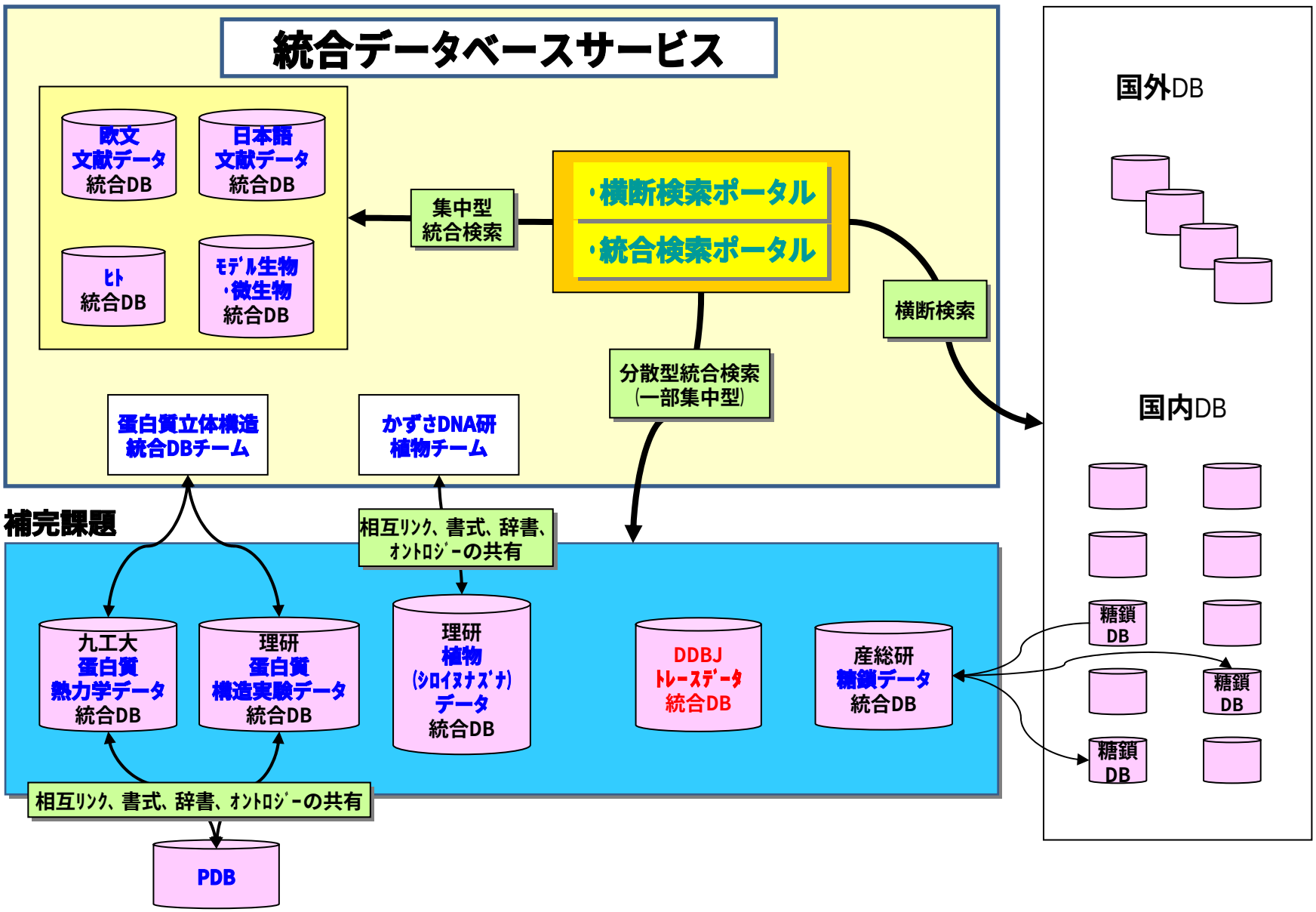
塩基配列アーカイブのデータベース構築と統合への貢献

事業の目標･概要

各種生物の遺伝子やゲノムの塩基配列を決定するいわゆるシーケンシングセンターにおいてはその配列決定の原データになる波形データなどのいわゆるアーカイブ(Trace archive; 以下Traceデータという)を有している。このTraceデータは、品質管理やそれを基にして行う配列決定アルゴリズムの改良ならびに配列断片を連結するアセンブリにおいて大変重要で貴重な情報である。そして、これらのTraceデータは、シーケンシングセンターの活動がその支持母体のプロジェクトの完了とともに終了するとすると、原則的には完全に消滅してしまう可能性が極めて高い状況にある。また、454やSolexaあるいはABI-SOLiDといった次世代の超高速の塩基配列決定装置の登場により、そのTraceデータの量は飛躍的に巨大化してきており、シーケンシングセンター自身でもそのデータハンドリングを含めて保存はもちろんのこと対処が非常に困難な状況になっている。

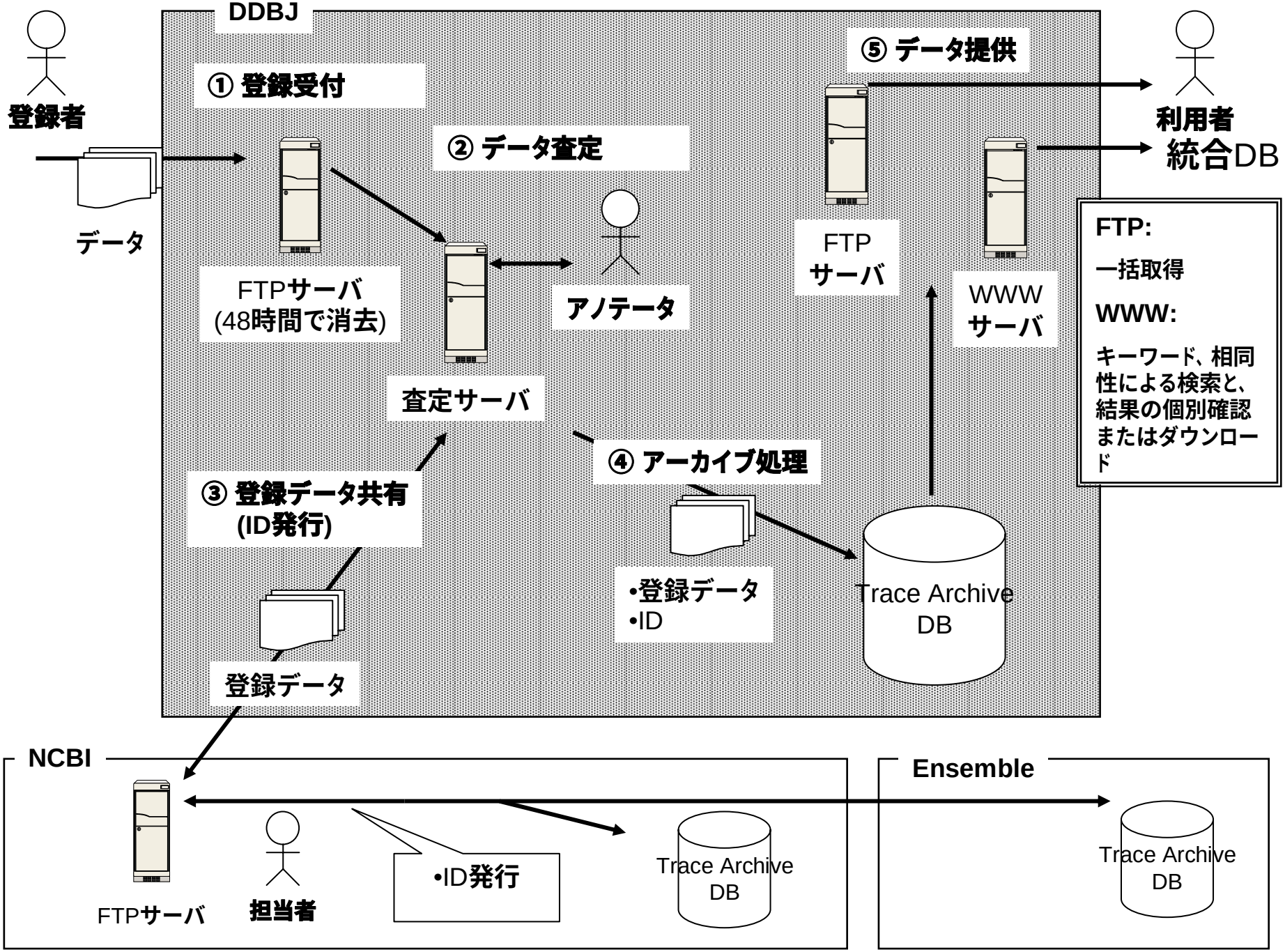
この状況の理解の下、わが国における塩基配列決定におけるTraceデータの保存と有効利用を目的として、当機関である大学共同利用機関法人　情報･システム研究機構の国立遺伝学研究所生命情報･ＤＤＢＪ研究センターのＤＤＢＪが、Traceデータのデータベース構築事業とデータ提供の事業を実施する。

補完課題における連携の概要

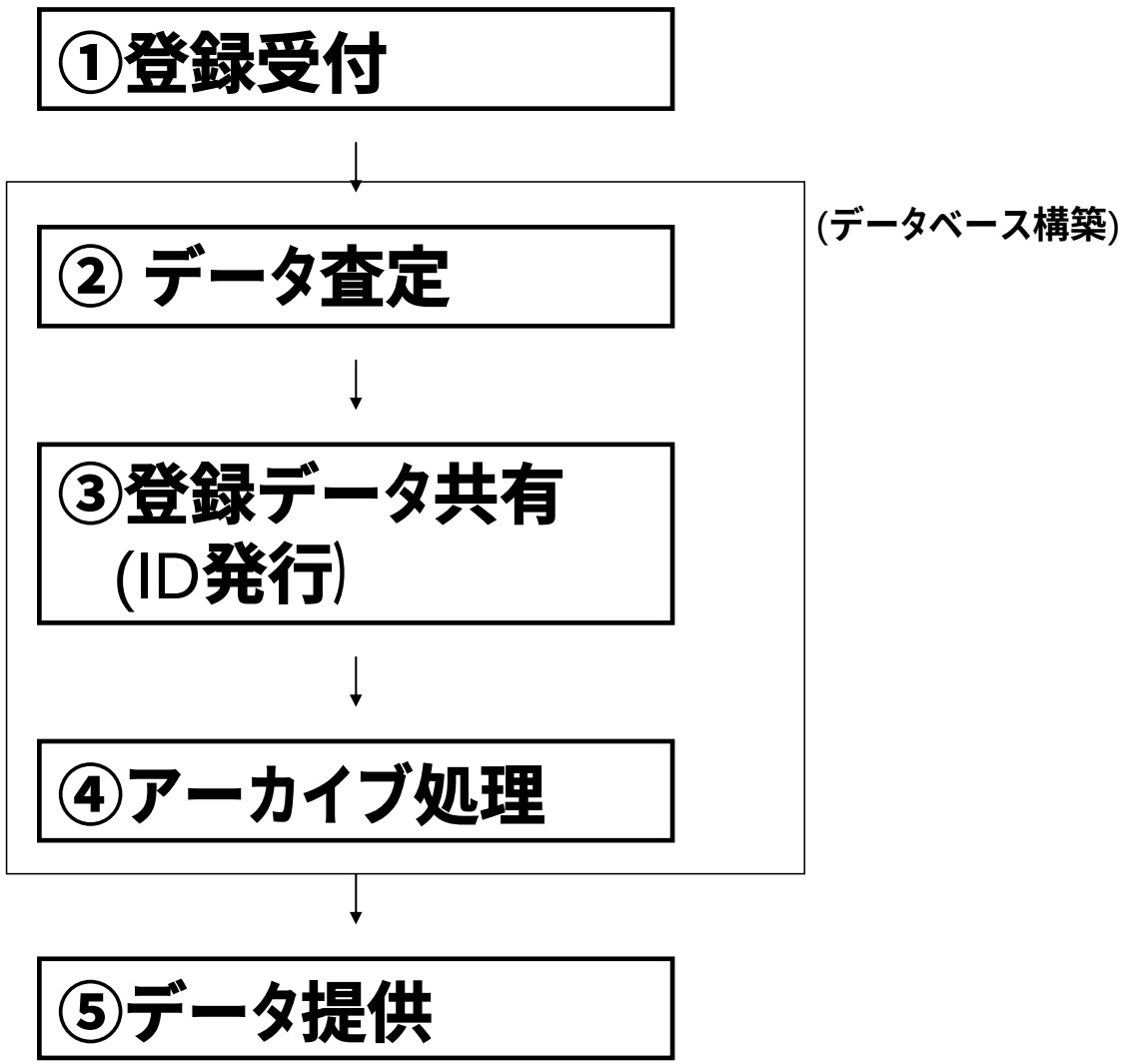


全体構成


登録者
遺伝研
遺伝研
東大
理研
かずさDNA
等
国内
シークエンス
センタ
①登録受付
②データ査定
③ID発行
④アーカイブ処理
⑤データ提供
FTP
Web
統合DB
利用者


DDBJ
登録者
① 登録受付
データ
FTPサーバ
(48時間で消去)
② データ査定
アノテータ
査定サーバ
③ 登録データ共有
(ID発行)
登録データ
④ アーカイブ処理
•登録データ
•ID
Trace Archive
DB
⑤ データ提供
FTP
サーバ
WWW
サーバ
利用者
統合DB
FTP:
一括取得
WWW:
キーワード、相同性による検索と、結果の個別確認またはダウンロード
NCBI
FTPサーバ
担当者
•ID発行
Trace Archive
DB
Ensemble
Trace Archive
DB

業務フロー



予測されるデータ規模の例

表１　アカデミアDNAシーケンシングセンターにおける解析中データ（遺伝研・小原所長）

					波形データ総容量 (Gbytes)			(2007年８月現在)
		リード数	プレート換算	HQリード数	非圧縮	圧縮(1)	圧縮(2)	
原始紅藻	GSZW	530,607	693	427,705	175	80	16	ゲノム（完成配列）登録済み
カタユウレイボヤ	GCiW	3,198,499	4,176	2,694,372	1,056	480	96	NCBI TraceArchive登録済み
メダカ	GOLW	15,675,095	20,464	12,925,557	5,173	2,351	470	ゲノム（ドラフト）登録済み
メダカ	GOLN	4,522,752	5,904	3,781,645	1,493	678	136	
マウス	GMS	10,249,629	13,381	9,173,139	3,382	1,537	307	
立襟鞭毛虫	GMOW	1,452,431	1,896	1,181,375	479	218	44	
細胞性粘菌	GASW	482,780	630	387,328	159	72	14	
Diploscapter	GNDW	1,089,826	1,423	809,113	360	163	33	
2006年度計		5,603,961	7,316	4,499,702	1,849	841	168	

注意：

・プレート換算は、リード数/766 で計算　(384ウェル - マーカ) x 表裏

・HQリード数は、HQ長>=300を閾値に算出

・波形データ容量は、ABI generic形式の非圧縮ファイル1件330Kbytesとして算出

・圧縮(1)は、gzip圧縮した結果を基に算出（1ファイル150Kbytes）

・圧縮(2)は、NCBI TraceArchiveにおける圧縮後SCFファイルのサイズより算出（1ファイル30Kbytes）

New sequencing technologies are drivers for new projects

- **454 Statistics**
- 100 Mb/run
- 250 base reads
- 400k reads/run
- 100x AB3730/day
- $12k/run
  - 20% AB3730 cost
- Homopolymer runs errors

- **Solexa Statistics**
- 8x 100 Mb/run
- 35 base reads
- 20 million reads/run
- 200x AB3730/day
- $5k/run
  - 1% AB3730 cost
- Diverse errors mainly base subn.

サービス	方式	資源	容量
Trace提供	・ データはディスク上に保管。 ・ テープ装置でバックアップ。 ・ WWWでの検索によるリクエスト受け付け。 ・ 一度の取得可能件数に制限を設ける。 ・ 結果はFTP用ディスクへ一時的に展開し、ファイル名等をメールで通知する。 ・ FTPデータは数日程度で削除する（取得完了したら削除する仕組みが望ましい）。	データ保管用ディスク テープ装置	それぞれ240TB以上
		FTP公開用ディスク	1TB 仮定： ・ 100万Trace = 50GB程度 ・ 100万Traceの要求が週に最大20件 ・ データは1週間の保存
FTP	・ Trace自体は置かない。（NCBI/Ensembl同様） ・ プロジェクト毎に、一定サイズごとに分割・圧縮したファイルを提供。	FTP公開用ディスク	11TB以上
キーワード検索	・ WWWからの検索リクエストを受け付け。 ・ 付随情報をDBに格納しておき、検索を実行する。	DB用ディスク DBサーバ	～4TB 仮定： インデックス等の領域をデータと同量 作業領域データと同量（最低線）
相同性検索	・ WWWからの検索リクエストを受付。 ・ BLASTなどのツールでの検索実行 ・ あらかじめ、配列からのインデクシングが必要。	内部用ディスク 検索サーバ	～6.2TB 仮定： ・ BLASTのDBを元の1/3量
WWW全般	・ キーワード及び相同性による検索機能。 ・ WWW上での検索結果表示では、波形も表示可能とする（要検討：十分なディスクを確保可能か）。 ・ WWW上での検索結果表示では、配列やクオリティスコアなどの表示も可能とする（要検討：ディスク容量）。 ・ FTPへのリンクや、Traceダウンロードのリクエスト機能を設置。	コンテンツ用ディスク ウェブアプリケーションサーバ	
波形表示	・ WWW上での即時の表示を提供する。 ・ ダウンロードしたデータを扱うために、波形表示用のプログラムをダウンロード提供する。 ・ 独自開発は行なわず、NCBIからのソースコード提供を受ける。		
登録処理	・ 登録データの受け付けは、メールでの連絡の後にFTP（またはセキュアFTP）により行なう。 ・ 受領したデータは、NCBIのTrace Archiveに登録処理を取り次ぐ（Ensmebl同様の方針）。 ・ ただし、登録元サイトの情報は、DDBJ内でも管理する。 ・ 登録者やNCBIとのやり取りは、構築局の作業を想定。支援的な半自動化システムは開発するが、完全自動化は行なわない。 ・ NCBIでの登録処理の結果は、自動的に取得し登録者用のFTPディレクトリに展開する。 ・ 登録処理が完了した元データは、一定期間後に削除する。		500GB 仮定： ・ 1登録当たり、大規模な場合200万Trace ・ 200万Trace = 100GB程度 ・ NCBIの登録処理に平均1週間程度と想定 ・ 週に最大5登録

試験システムの開発

実施項目	１年度目	２年度目	３年度目	４年度目
（１）公開FTPサイトとWWWサイトに関する開発 （参画研究機関） なし	**基盤技術整備**	公開サイトシステムの整備と開発		取りまとめ
（２）登録処理および波形表示システムに関する開発 （参画研究機関） なし	**試験システムの開発**	本システムの整備と開発		