出版界の外字・異体字問題を考える

「字形共通基盤」プロトタイプによる実証実験のご紹介

経済産業省委託事業「平成22年度書籍等デジタル化推進事業」

（デジタル・ネットワーク社会における出版物の利活用推進のための外字・異体字利用環境整備事業）

2011年12月

凸版印刷株式会社

アジェンダ

1. 全体概要
2. 字形共通基盤のプロトタイプ
3. 実証実験の実施要項
4. Q&A

1.

全体概要

1. 経緯

step	経緯	ポイント
1	三省デジ懇 2010年8月	外字・異体字問題に関しても解決が必要と提起
2	調査検討 2011年1月～3月	外字・異体字の理想的な利用方法に関し、調査検討を行い、進むべき方向性を定める
3	実証実験 2011年4月～2012年2月	調査検討で定めた方向性を実証実験により検証し、並行して運用課題も検討する

2. 調査検討フェーズの調査項目

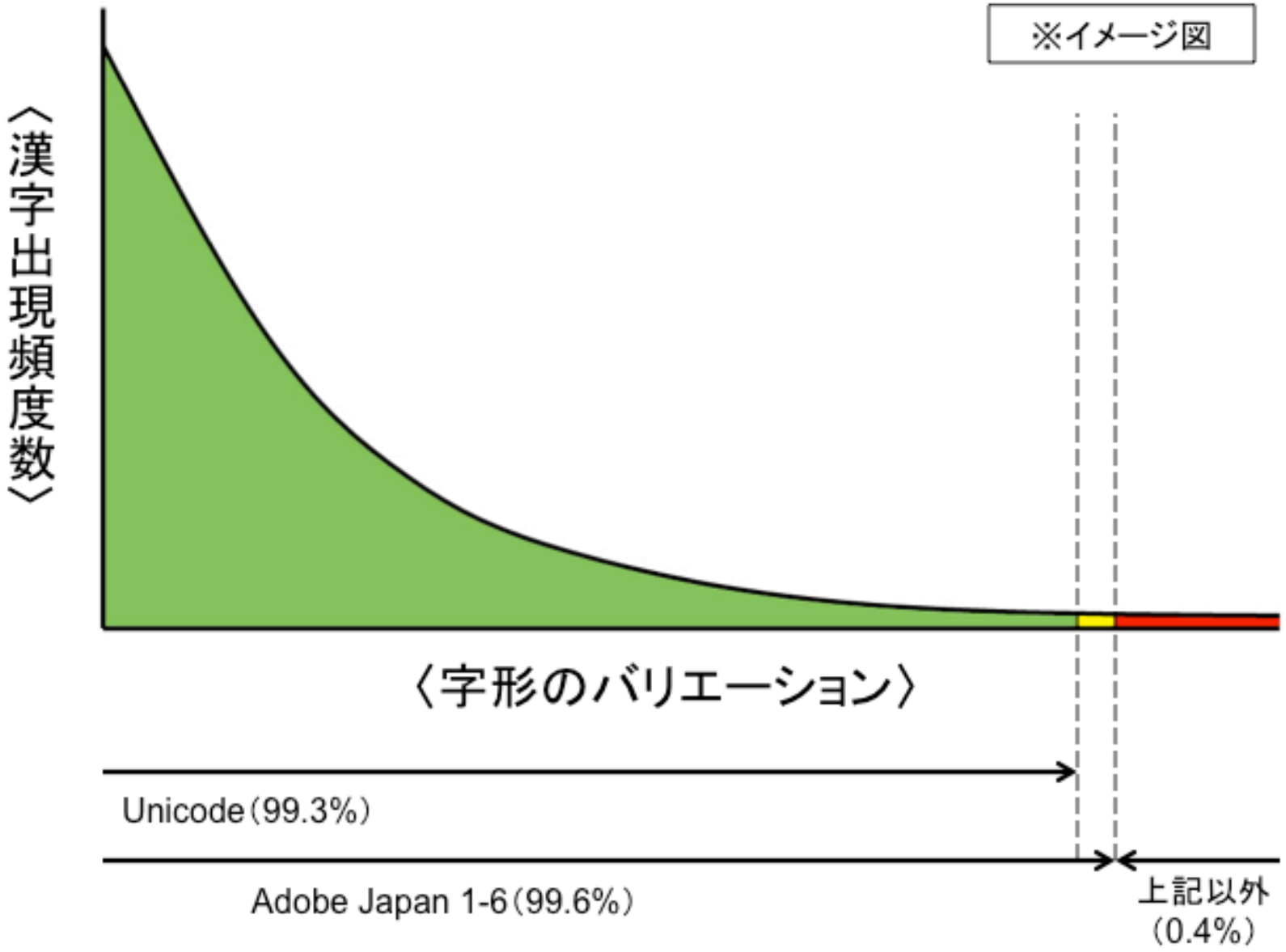
ID	調査項目	概略
1	文字鏡研究会	諸国の文化を支えていた文字を利用可能にする調査研究。文字番号の採番とフォントの配布
2	インデックスフォント研究会	コード表にない漢字等へユニークな番号を付与し、対応フォント等を整備。新聞、出版印刷、BF、官公庁等をターゲット
3	GT明朝（TRONプロジェクト）	ユビキタス社会において、誰でも扱えるTRON多言語環境を実現。GT明朝は、その漢字面の一部
4	CHISE	文字コードを使わず文字処理が行える状況を確立させ、符号化文字集合に含まれない文字も容易に扱える
5	漢字データベース	UCS（CJK統合漢字）を使い易くし、利用促進させるDB
6	グリフウィキ	文字の「青天井問題」に対するソフトウェアによる解決
7	漢字出現頻度数調査	常用漢字改訂の基礎資料として調査された漢字出現頻度数調査を、出版物の外字・異体字視点から分析
8	「広辞苑」で使われている字形数	広辞苑（第六版）に使われている字形数の調査・分析
9	外字・異体字対応フロー	各社、いちど内部コードに変換して対応。外字が発生すれば作字し、各社独自の採番によって管理される
10	データ配信事業の外字・異体字	版元の意向に基づき、画像か内字に置き換え対応
11	文字情報基盤構築事業	行政処理の合理化を目指した文字情報基盤の構築事業

［参考-1］漢字出現頻度数調査

各種範囲			漢字数		字形数	
Adobe Japan 1-6	Unicode	JIS X 0208	47,542,535	99.2%	5,774	67.3%
		JIS X 0208以外	70,049	0.1%	1,426	16.6%
	CIDのみ		140,028	0.3%	393	4.6%
上記以外			171,821	0.4%	983	11.5%
合計			47,924,433	100.0%	8,576	100.0%

- 凸版印刷が自社のCTSデータ（800冊分、出現漢字数5,000万字弱）を使って、2007年に行った漢字出現頻度数調査の結果で、出現頻度の低い漢字に注目
- 出現した漢字の99.6%は、Adobe Japan 1-6に包含されていた

［参考-1］漢字出現頻度数調査



［参考-2］広辞苑で使われている字形数

岩波書店「広辞苑」第六版

範囲	字形数（概算）	
1）JIS第1水準、第2水準	6,355	34.1%
2）補助漢字（JIS X 0212）	5,801	31.1%
3）Unicode（JIS X 0221）	5,300	28.4%
4）ユーザ外字	1,200	6.4%
合計	18,656	100.0%

JIS第3水準、第4水準の字種は、上記の2）―4）の中に概ね包含されている

3. 問題点の切り分け

区分	作り手側		読者側
工程	執筆・編集	情報加工・蓄積	情報公開（出版）
特徴	知の創造活動	グリフの性質上、漢字の出現頻度数に関係なく、膨大な字形が存在する（ロングテール）	端末によって符号化文字集合の対応が異なり、内字／外字の状況が変わる
問題点	外字・異体字指示が直接行えない場合があり、ゲラでのやりとり（赤字指示）が無くならない	外字・異体字判定やデータ化方式がバラバラで、互換性を保てないリスクが高く、対応コストも高い	外字・異体字を正確に表示できない（又は検索できない）場合がある
方向性	• 作業支援ツールの整備	• 字形判定・格納基盤の整備 • 例示フォントの整備 • 漢字属性情報の整備 • 作業支援ツールの整備	• 国際標準規格の利用推進 • 外字表現方法の整備 • 書体の拡充 • 利用者支援ツールの整備

4. 区別が求められるグリフとコンピュータの関係

出版物

出版界で一般的に区別が求められるグリフ

（こだわり、嗜好等を含む）

（A）

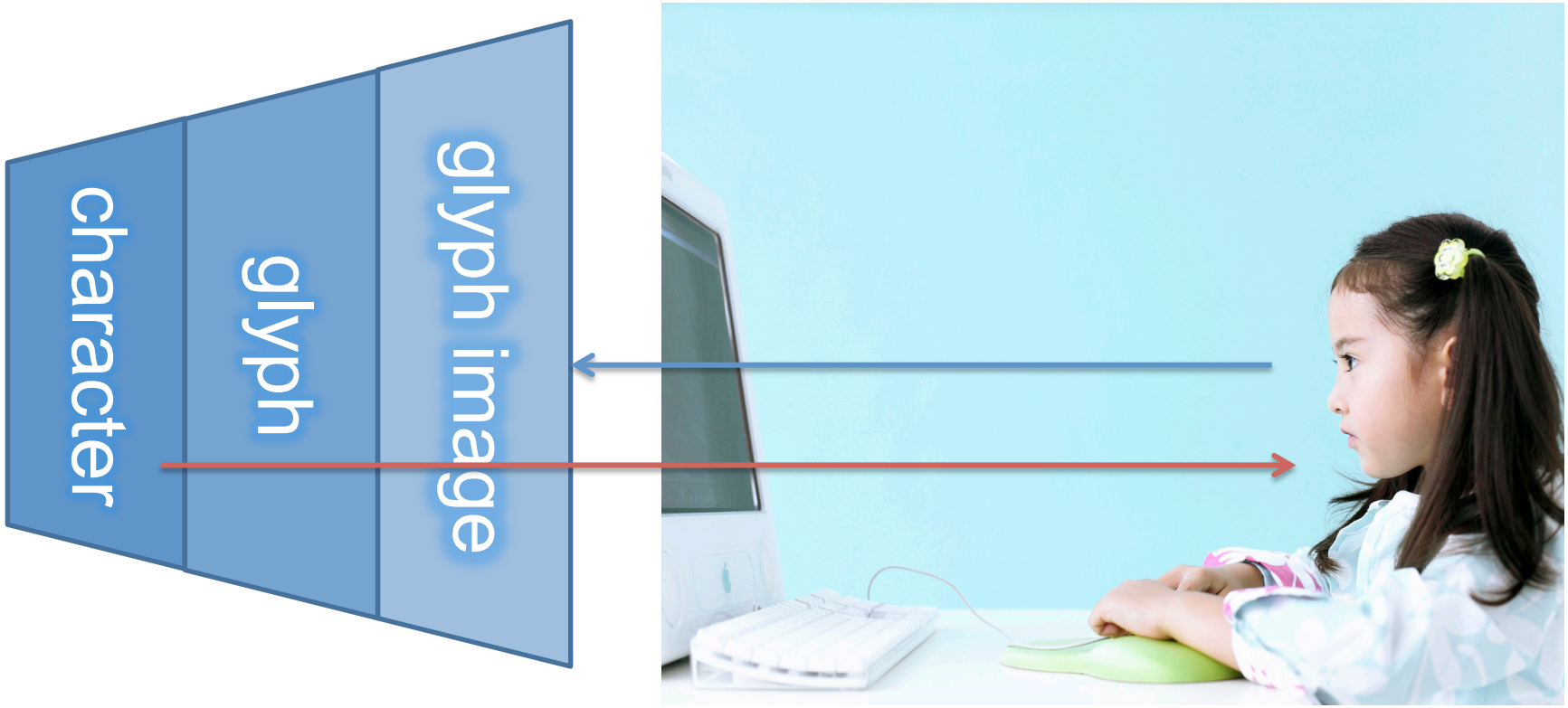
＞

コンピュータ

符号化文字集合の符号化された例示字形によるフォント

（B）

［参考-3］文字の視覚的表現におけるレイヤー概念



［参考-4］この資料における用語の定義

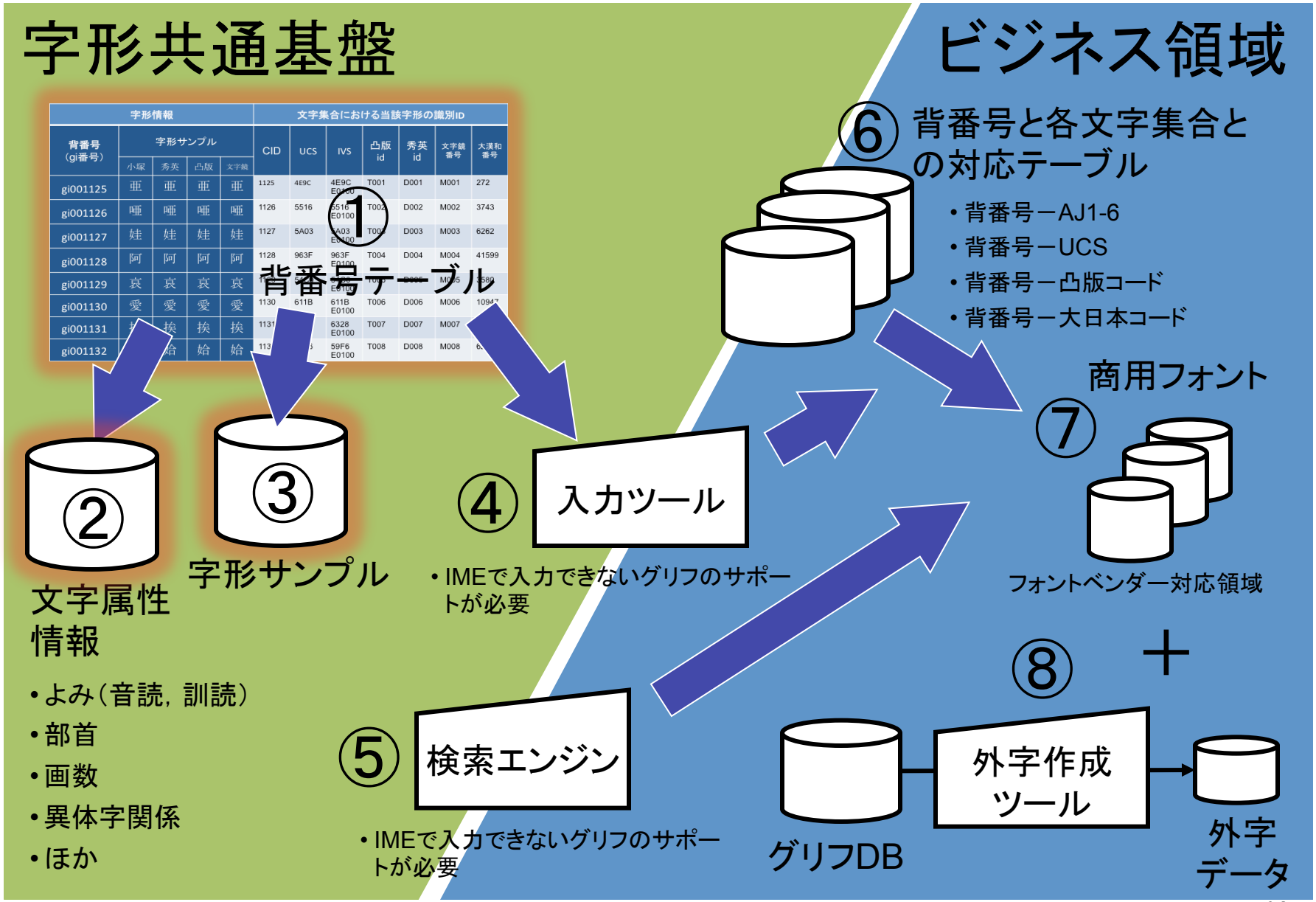
文字（character）	言語において意味をもつ最小単位。特定の形状のことをさす物ではなく、抽象的な意味と抽象的な形状のことを指す。
グリフ（glyph）	グリフイメージを表現する抽象形式（abstract shape）
グリフイメージ（glyph image）	何らかの表示媒体（コンピュータディスプレイや紙など）の上に描いた、グリフ表現の具体的な画像
フォント（font）	文字の視覚的表現のために使われるグリフを集めたもの

5. 外字・異体字対応の方向性

- 出版界で一般的に区別することが求められるグリフを収集整理して、共通インフラとして構築する（字形共通基盤）

- 具体的には、対象となるグリフに識別ID（背番号）を付与してデータベース化し、管理運用する

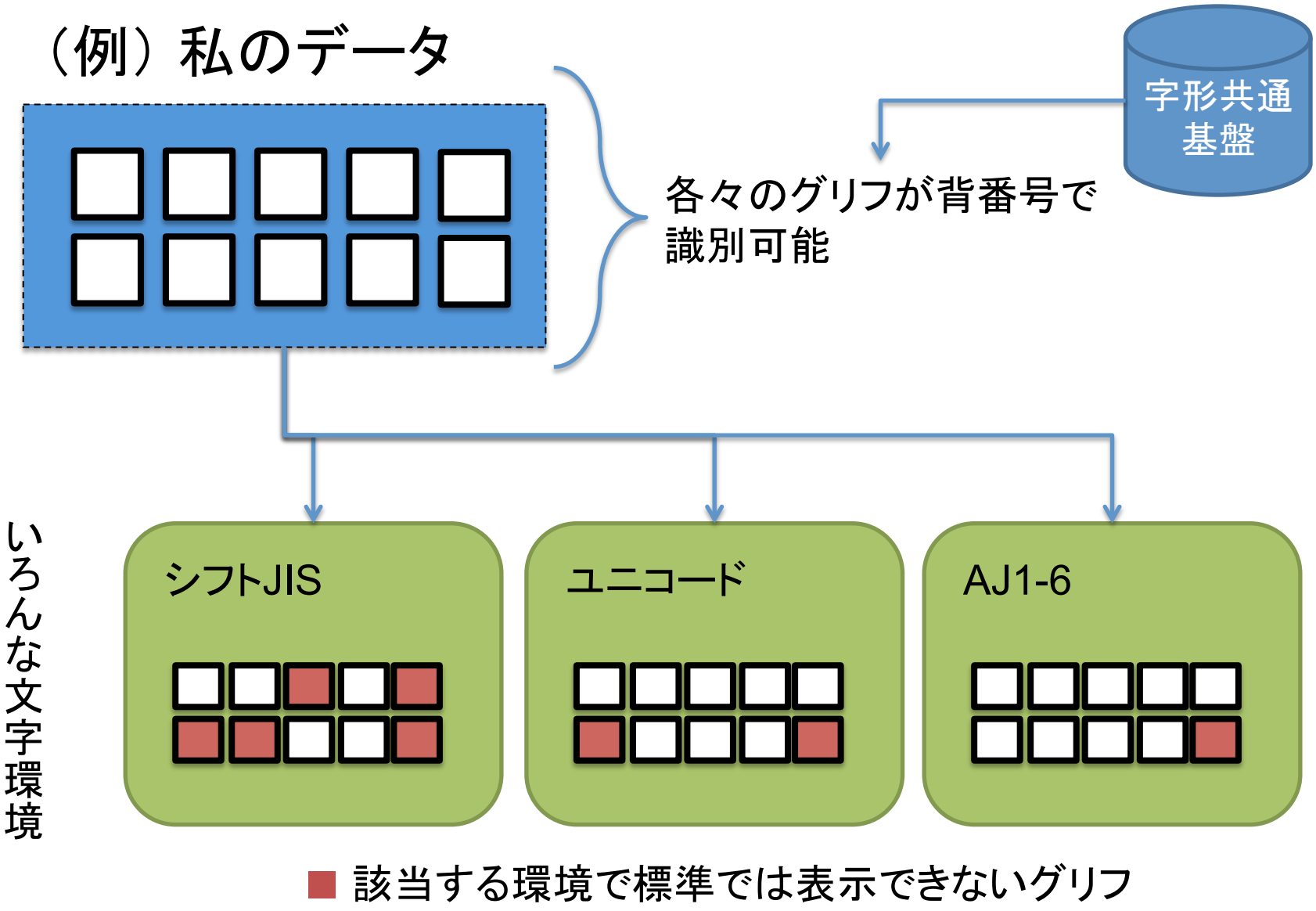
6. 想定するフロー


字形共通基盤
ビジネス領域
字形情報
文字集合における当該字形の識別ID
背番号（gi番号）
字形サンプル
小塚
秀英
凸版
文字鏡
CID
UCS
IVS
凸版id
秀英id
文字鏡番号
大漢和番号
gi001125
gi001126
gi001127
gi001128
gi001129
gi001130
gi001131
gi001132
①
背番号テーブル
⑥
背番号と各文字集合との対応テーブル
• 背番号－AJ1-6
• 背番号－UCS
• 背番号－凸版コード
• 背番号－大日本コード
②
文字属性情報
• よみ（音読，訓読）
• 部首
• 画数
• 異体字関係
• ほか
③
字形サンプル
④
入力ツール
• IMEで入力できないグリフのサポートが必要
商用フォント
⑦
フォントベンダー対応領域
⑤
検索エンジン
• IMEで入力できないグリフのサポートが必要
⑧
＋
グリフDB
外字作成ツール
外字データ

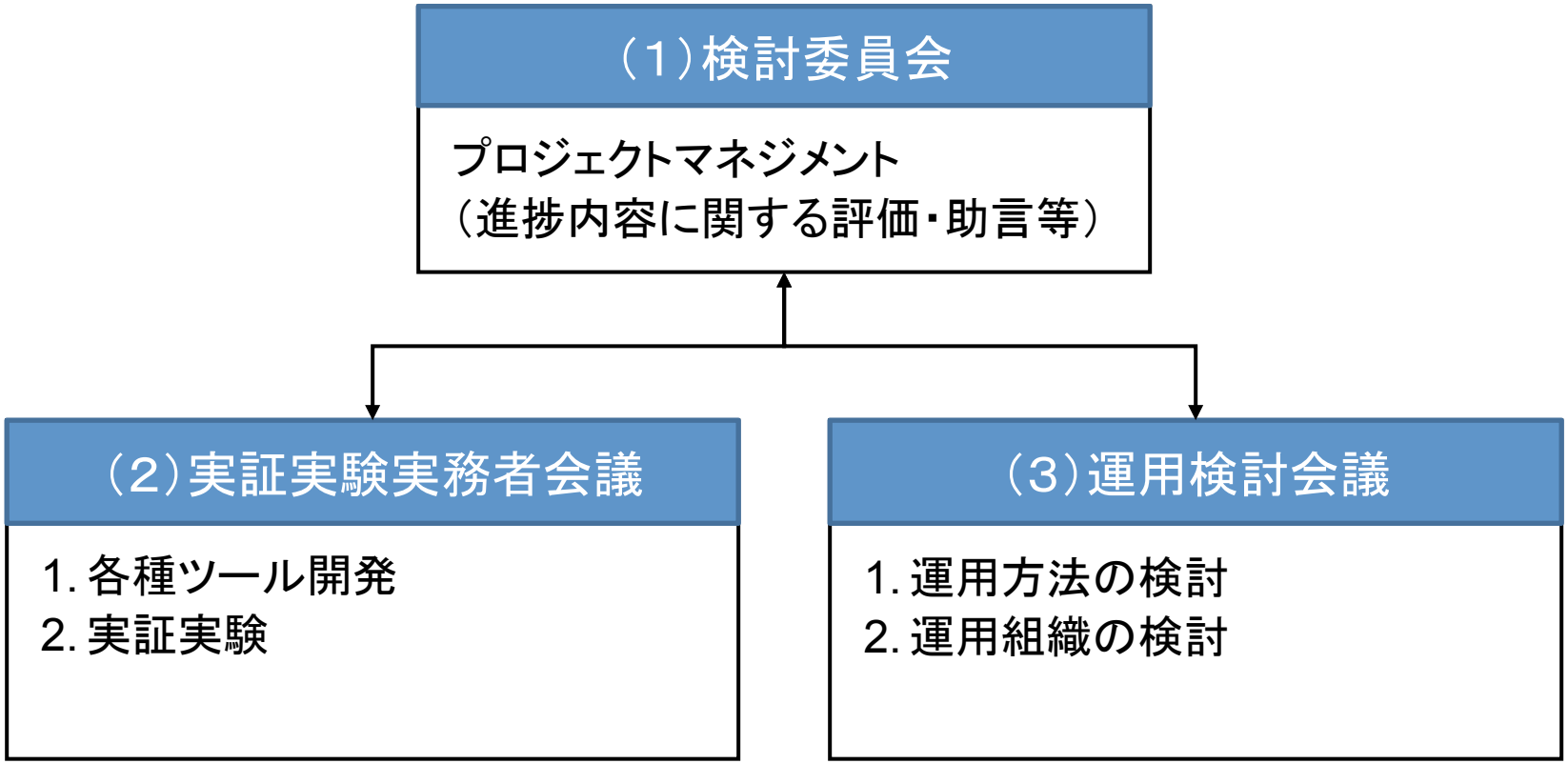
7. 背番号テーブルのイメージ

字形情報					文字集合における当該字形の識別ID						
背番号（gi番号）	字形サンプル				CID	UCS	IVS	凸版id	秀英id	文字鏡番号	大漢和番号
	小塚	秀英	凸版	文字鏡							
gi001125	亜	亜	亜	亜	1125	4E9C	4E9C E0100	T001	D001	M001	272
gi001126	唖	唖	唖	唖	1126	5516	5516 E0100	T002	D002	M002	3743
gi001127	娃	娃	娃	娃	1127	5A03	5A03 E0100	T003	D003	M003	6262
gi001128	阿	阿	阿	阿	1128	963F	963F E0100	T004	D004	M004	41599
gi001129	哀	哀	哀	哀	1129	54C0	54C0 E0100	T005	D005	M005	3580
gi001130	愛	愛	愛	愛	1130	611B	611B E0100	T006	D006	M006	10947
gi001131	挨	挨	挨	挨	1131	6328	6328 E0100	T007	D007	M007	12082
gi001132	姶	姶	姶	姶	1132	59F6	59F6 E0100	T008	D008	M008	6242

8. 字形共通基盤と、いろんな文字環境との関係



9. 実証実験フェーズの活動概要



- 実験結果（有用性評価結果）
- 背番号テーブル、字形サンプル、漢字属性（ミニマムセット）
- 入力ツールと検索ツールのプロト版
- 残された課題

- 運用方法案（受入ルール、登録フロー）
- 運用組織案

10. 検討委員会（ミッション：プロジェクトマネジメント）

座長	三田　誠広	作家	公益社団法人　日本文藝家協会　副理事長
副座長	小林　龍生	Unicode	Unicode Consortium　Director
委員	相田　満	有識者	大学共同利用機関法人　人間文化研究機構　国文学研究資料館・研究部　准教授
	長村　玄	有識者	インデックスフォント研究会　幹事会顧問
	黒田　信二郎	JEPA	一般社団法人　日本電子出版協会　文字図形共有基盤調査検討分科会　委員長（紀伊国屋書店）
	新名　新	電書協	一般社団法人　日本電子書籍出版社協会　常任理事（角川書店）
	平井　彰司	書協	社団法人　日本書籍出版協会　知的財産権委員会　副委員長（筑摩書房）
	丸山　信人	雑協	社団法人　日本雑誌協会　デジタルコンテンツ推進委員会 幹事（インプレスホールディングス）
	植村　八潮	出版	日本出版学会　副会長（東京電機大学出版局）
	冨田　信雄	フォント	株式会社モリサワ　デジタルタイプセンター　部長
	三橋　洋一	フォント	大日本スクリーン製造株式会社　メディア＆プレシジョンテクノロジーカンパニー
	山本　太郎	ソフト	アドビシステムズ株式会社　エンジニアリング　シニア・マネージャー
	加治佐　俊一	ソフト	日本マイクロソフト株式会社　業務執行役員　最高技術責任者
	堀口　宗男	印刷業界	社団法人　日本印刷産業連合会
	千葉　弘幸	印刷業界	社団法人　日本印刷技術協会
オブザーバ	亀井　義人	印刷	凸版印刷株式会社　製造・技術・研究本部 部長
	高橋　仁一	印刷	大日本印刷株式会社　C&I事業部 IT開発本部 秀英体開発室　室長
	高柳　大輔	官庁	経済産業省　商務情報政策局　文化情報関連産業課（メディア・コンテンツ課）課長補佐
	松田　昇剛	官庁	総務省　情報流通行政局　情報流通進行課　統括補佐

（2011.5.10）

11. 実証実験実務者会議（ミッション：実証実験の実施）

座長	田原　恭二	凸版印刷	実務者会議PM
副座長	高橋　仁一	大日本印刷	実務者会議PM
委員	秋元　良仁	凸版印刷	実務作業全般
	宮田　愛子	大日本印刷	背番号テーブル該当情報抽出・セット
	喜多　英司	ジャストシステム	クラウド型　入力ツール、検索ツールプロトタイプ開発
	田中　和広	ジャストシステム	クラウド型　入力ツール、検索ツールプロトタイプ開発
	岩田　真一	Indexfont研究会	背番号テーブル該当情報抽出
	上地　宏一	大東文化大学	漢字データベース、その他CHISE情報、グリフウィキ情報等の提供
	福島　慎太郎	出版	技術コメント（出版社視点）
	小池　利明	ボイジャー	技術コメント（電子書籍視点）
	斎鹿　尚史	シャープ	技術コメント（電子書籍視点）
	石井　宏治	W3C/CSS Editor	技術コメント（Web技術視点）
	増田　浩一	モリサワ	技術コメント（フォント視点）
オブザーバ	小林　龍生	検討委員会副座長	技術アドバイス

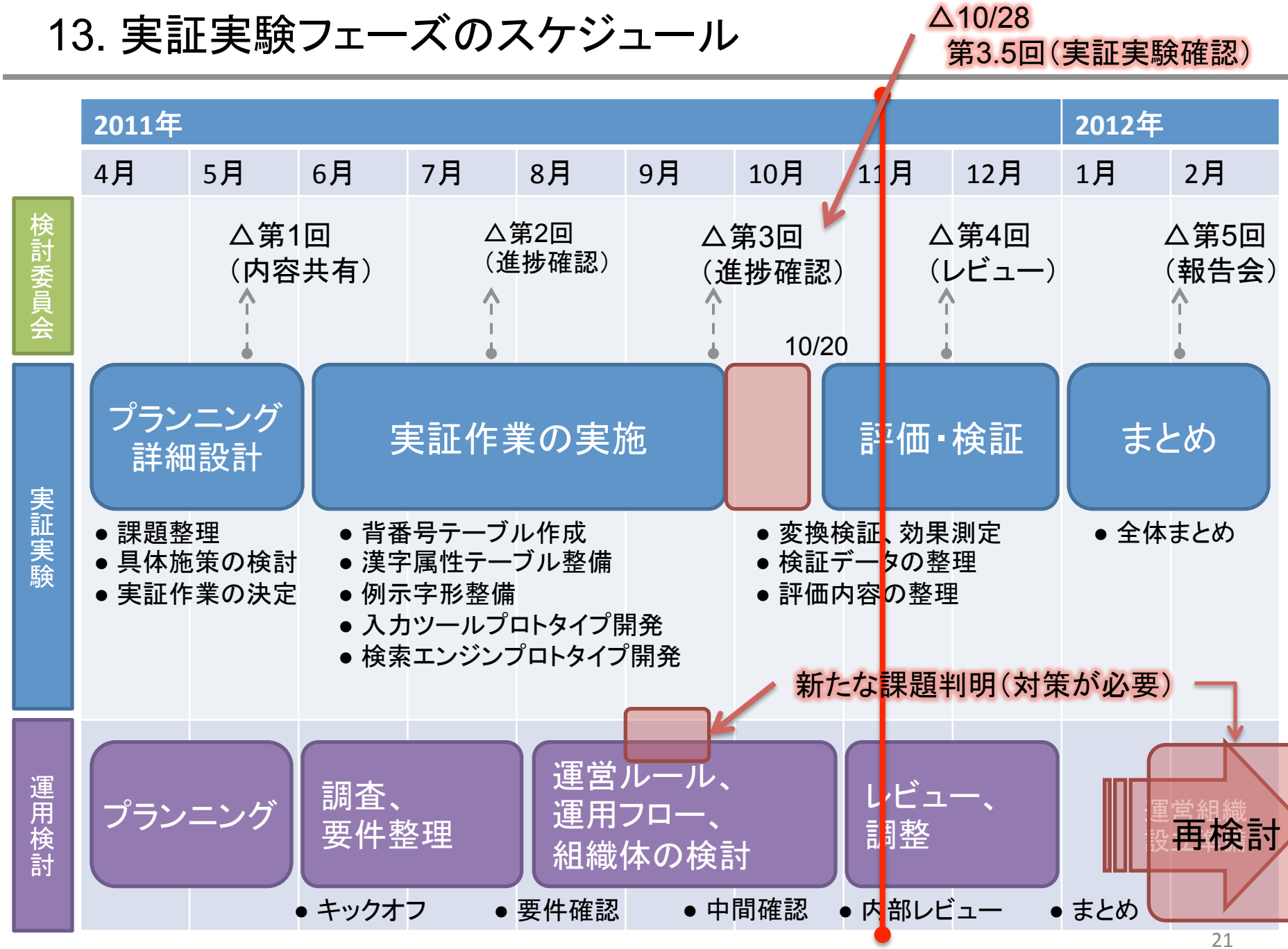
（2011.7.26）

12. 運用検討会議（ミッション：運用課題の検討）

座長	植村　八潮	日本出版学会	日本出版学会　副会長
副座長	高野　郁子	出版社	三省堂　出版局デジタル情報出版部部長
委員	田中　正明	出版社	岩波書店　編集局　部長（辞典編集部・電子出版部担当）
	長村　玄	indexfont	インデックスフォント研究会　幹事会顧問
	黒田　信二郎	JEPA	一般社団法人　日本電子出版協会　文字図形共有基盤調査検討分科会　委員長
	丸山　信人	出版社	インプレス・ホールディングス　執行役員
	鎌仲　宏治	印刷会社	凸版印刷株式会社　営業本部長
	福田　健一	印刷会社	大日本印刷株式会社　市谷事業部　副事業部長
	川崎　誠一	電子出版流通	一般社団法人　電子出版制作・流通協議会
	堀口　宗男	印刷業界	社団法人　日本印刷産業連合会
	千葉　弘幸	印刷業界	社団法人　日本印刷技術協会
	岡本　和之	印刷業界	印刷工業会　理事
オブザーバ	亀井　義人	印刷会社	凸版印刷株式会社
	高橋　仁一	印刷会社	大日本印刷株式会社
	小林　龍生	検討委員会副座長	
	高柳　大輔	官庁	経済産業省　商務情報政策局　文化情報関連産業課（メディア・コンテンツ課）

（2011.7.26）

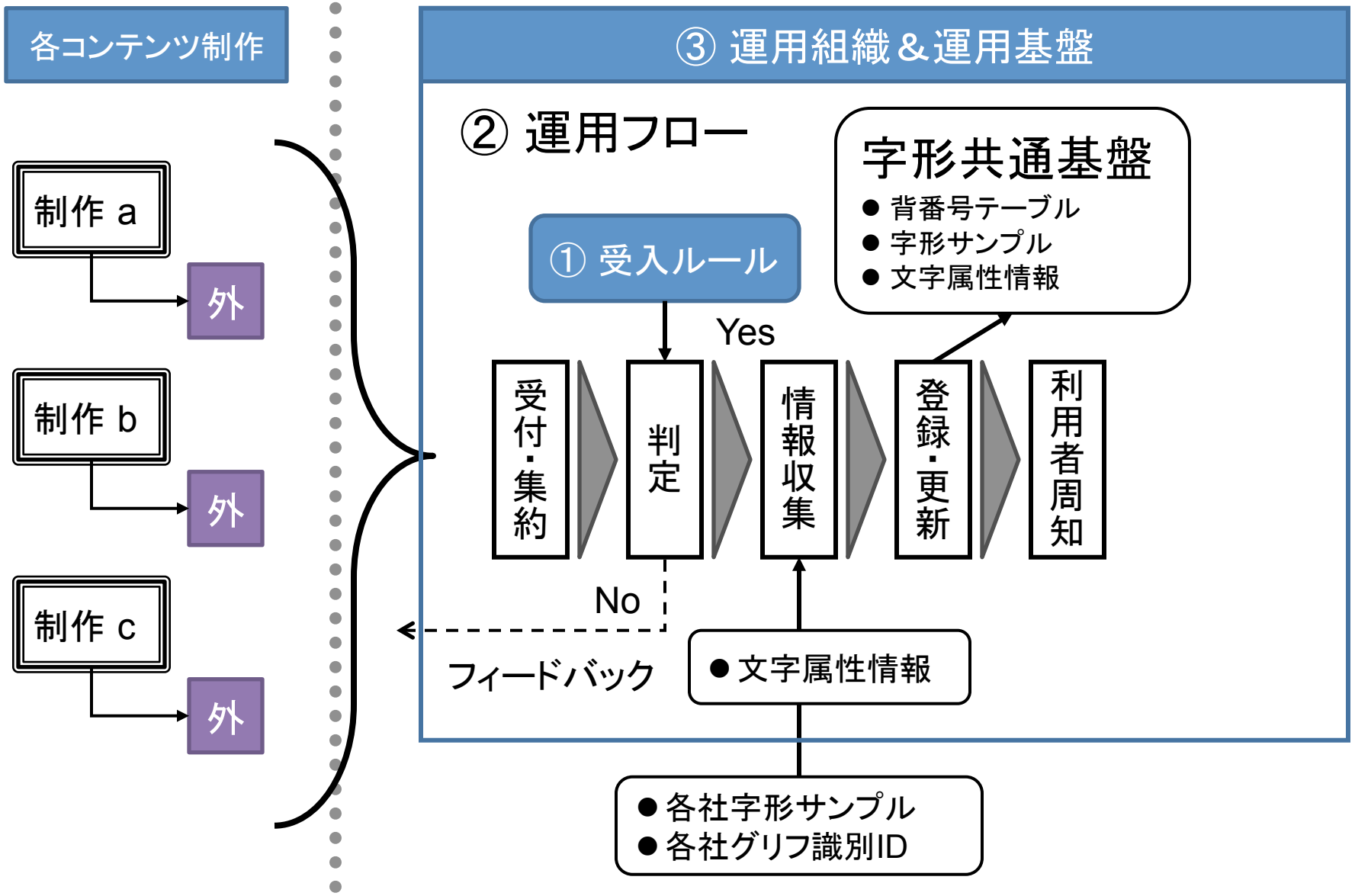
13. 実証実験フェーズのスケジュール



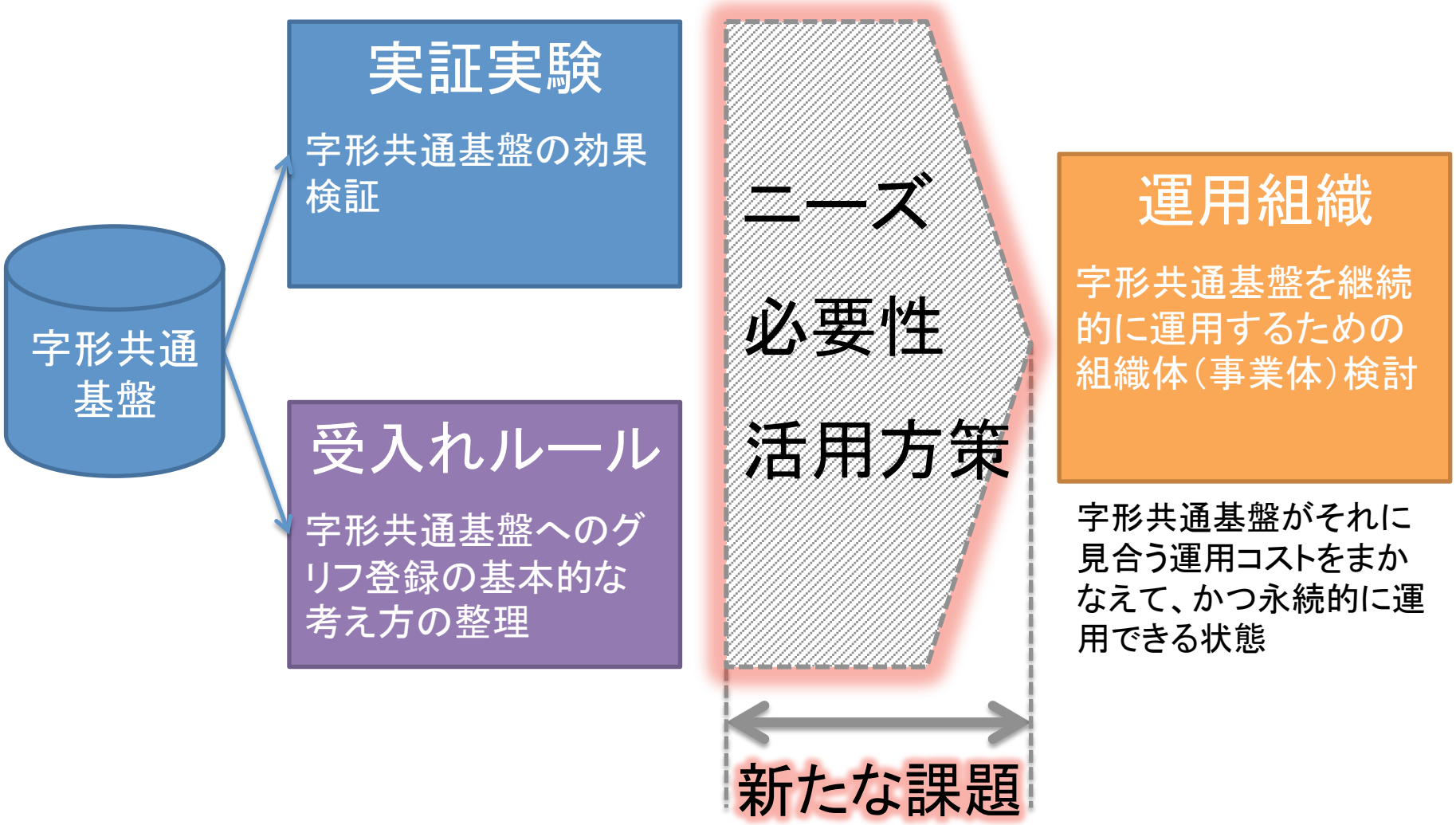
14. 現在の進捗状況

課題	状況
実証実験	● 10月末よりツールを配布し、実験開始
受入ルールの検討	● 運用検討会議サブグループで進行中 - IVD AJ1コレクションに登録されるグリフの特徴点を検証し、受入ルールを検討中 - サンプルコンテンツの外字を確認中 ● 12月に受入ルール案（ver.1）が完成し、2011年1月にレビュー
運用組織の検討	● 運用組織の検討より前に、明らかにすべき課題（必要性・具体的な活用方策の抽出）が判明 ● 実証実験を踏まえて、今後の進め方を含め、再度検討を行う（2011年1月予定）

15. 字形共通基盤の運用イメージ



16. 運用組織より前に、具体的に明らかにする必要がある

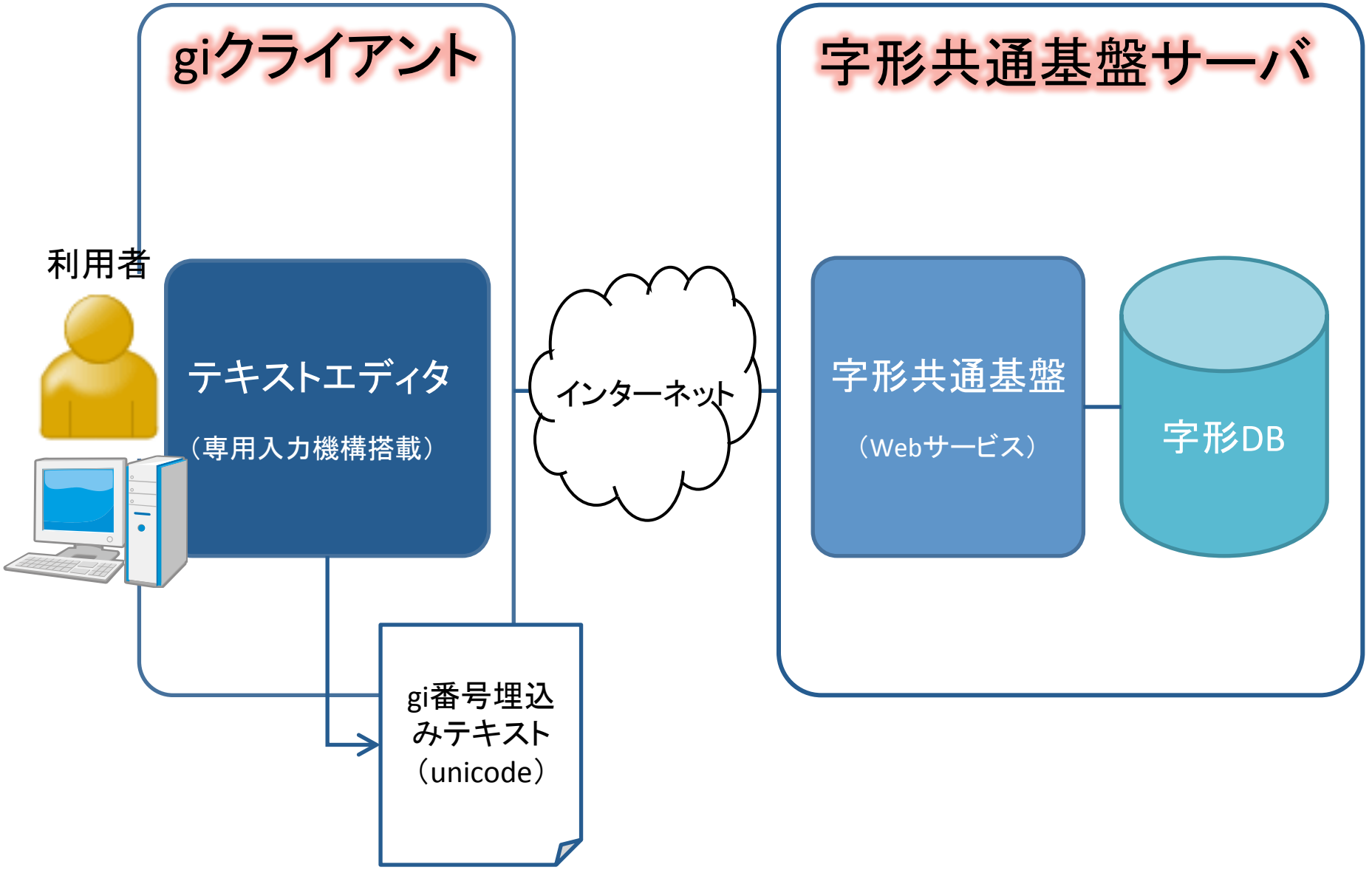


※ 把握不足・十分議論されていない

2.

字形共通基盤プロトタイプのご紹介

1. 概念図


giクライアント
字形共通基盤サーバ
利用者
テキストエディタ
（専用入力機構搭載）
インターネット
字形共通基盤
（Webサービス）
字形DB
gi番号埋込
みテキスト
（unicode）

2. 字形共通基盤プロトタイプのデモンストレーション

画面にご注目ください。

3. 字形共通基盤プロトタイプ

項目	内容
登録グリフ	実験スタート時のボリュームとしては、Adobe Japan 1-6のグリフセットを登録。 新たに出現したグリフは順次追加していく
字形サンプル	小塚明朝、秀英体、凸版明朝、文字鏡、 ヒラギノ、リューミン（調整中）
検索機能	文字属性情報を使って検索が可能 （なお、一部の属性情報はプロトタイプ版では対応していないものがあります）

4. 背番号の表記（gi番号形式）

- グリフを識別するid（背番号）として、次の形式による一意のidを割り当てる

プレフィックス（gi）　＋　数字6桁

- 数字6桁は整数（ゼロ埋めして表現）
- 背番号は永久欠番とする

5. 字形サンプルのスペック

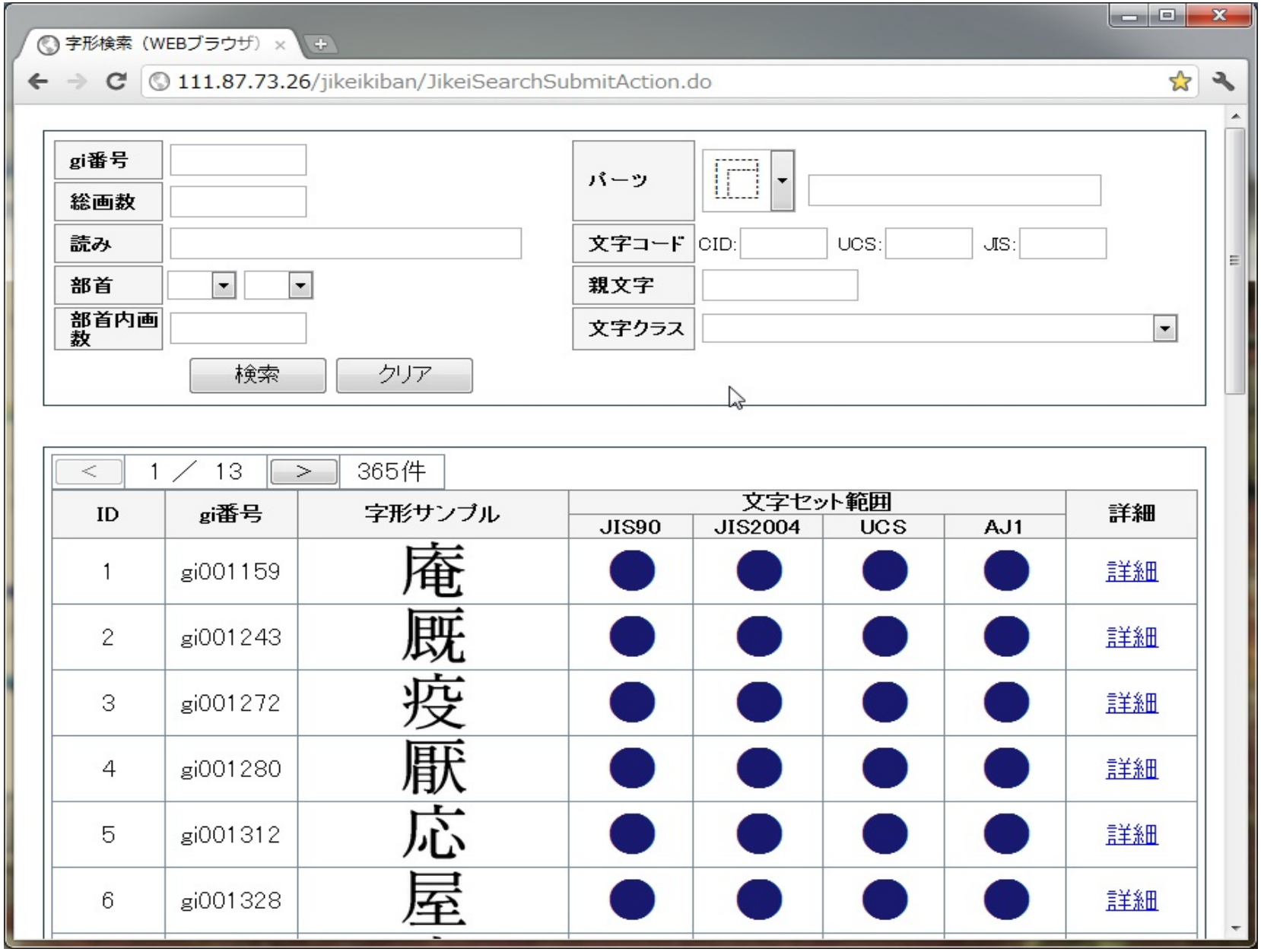
（例）

- 128×128pixel（PNG）
- color：black, background color: white

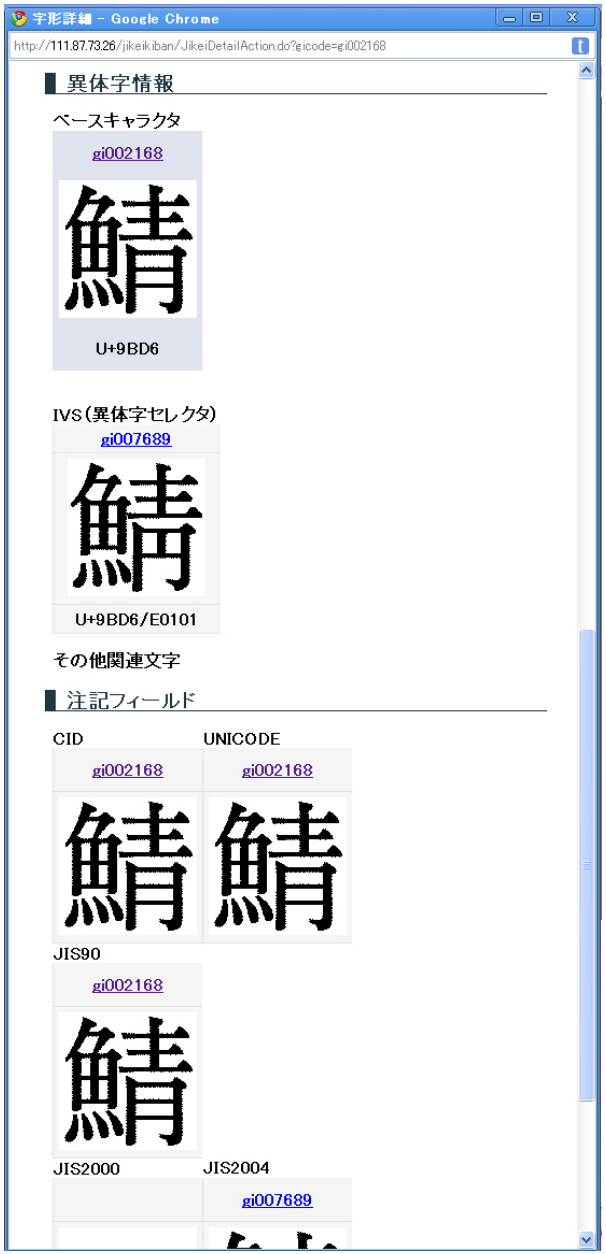
6. 各グリフの文字属性情報

部首、部首画数	字体変更情報
読み	康熙別掲字、CJK互換漢字、その他関連字
漢字構成記述文字	縦横区分
CID	文字クラス
UCS	代替文字
JIS	意味
IVD AJ1コレクション	登録者情報
大漢和番号	

7. 字形共通基盤サーバー／検索画面（例）



8. 字形共通基盤サーバー／グリフの詳細情報（例）

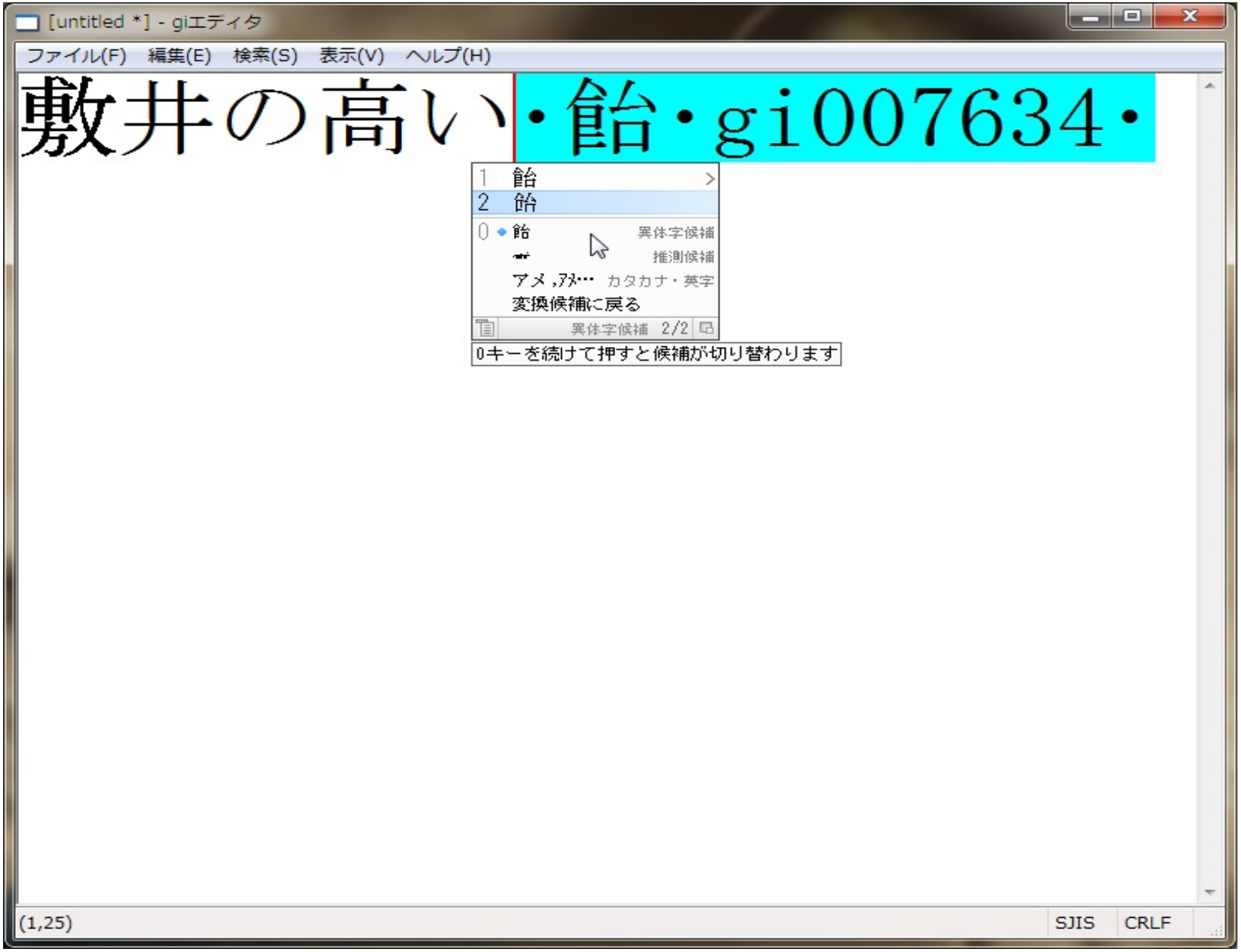


9. 字形共通基盤サーバー／スマートフォン（例）

10. giクライアント

項目	内容
テキストエディタ	字形共通基盤との通信機能を持ったシンプルなテキストエディタ
入力機能	字形共通基盤との通信機能をもった専用のATOK
保存形式	ユニコードテキストとして保存される 字形共通基盤から入力したグリフはUnicodeのInterlinear Annotationと同等の形式で保存される

11. 字形共通基盤との通信機能を持ったテキストエディタ



3.

実証実験の実施要項

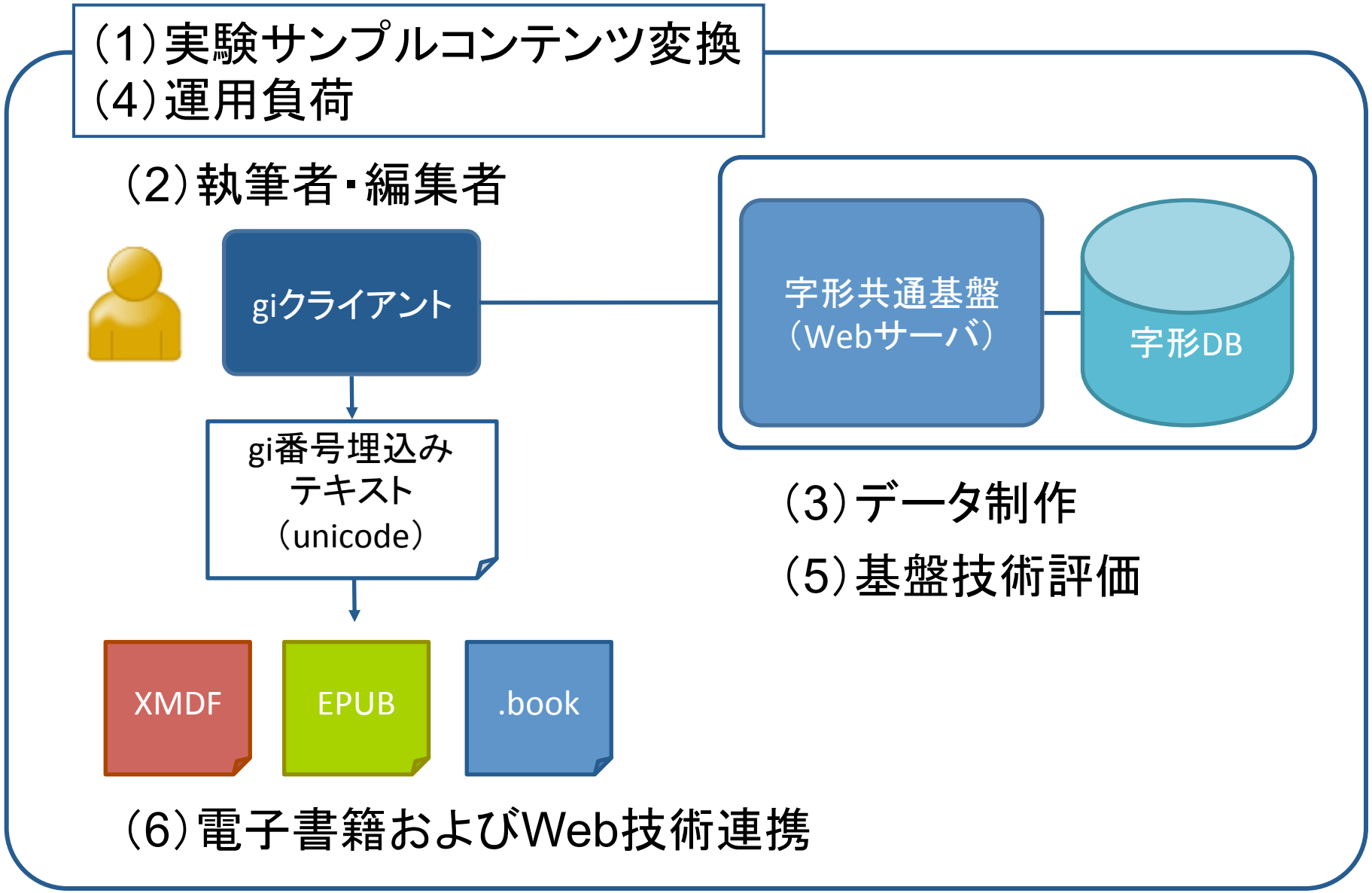
1. 実証実験の実施要項

項目	概要
実験期間	2011年10月28日～12月末日（約2ヶ月）
利用環境	● Windows XP以上（Macは利用不可） ● Internet Explore 8以上 ● giクライアント（インストールが必要） ● 字形共通基盤アクセスに専用の id/pwd が必要（id/pwdは事務局からメールにて個別にご案内いたします）
利用方法	Webサイトからgiクライアントインストーラと利用マニュアルをダウンロードして利用 ※疑問点などはMLを使ってフォローいたします。
実験のポイント	● 字形共通基盤を使った外字・異体字入力の確認 ● 字形共通基盤を使った外字・異体字表示の確認 ● 受入ルールの適合性検証と運用作業負荷の把握 ● 字形共通基盤の必要性の検証と活用方策の検討
その他	実験参加の同意書にご同意をお願いします。

2. 実証実験の分類とポイント

ID	分類	ポイント	想定する被験者
(1)	実験サンプルコンテンツ変換	サンプルコンテンツを使って、各出力メディアで、想定するグリフが表示できるか否かを確認	凸版印刷
(2)	執筆者・編集者	著者⇔編集者間の文字入力、編集作業等の負荷軽減確認	執筆者、編集者、校閲者
(3)	データ制作	字形基盤の必要性の確認 活用ニーズの抽出・把握	運用検討SWGメンバー、制作業務に関係する各協会、研究会など
(4)	運用負荷	サンプルコンテンツで出た外字・異体字を使って、受入ルール判定および運用負荷を把握	運用検討SWGメンバー
(5)	基盤技術評価	情報の正確さ システムの信頼性・可用性・保守性	実務者会議メンバー
(6)	電子書籍およびWeb技術連携	連携に向けた方向性の見極め	実務者会議メンバー

3. 実証実験の分類体系図



4. 字形共通基盤プロトタイプのご紹介スケジュール

ID	委員／団体	状況／予定
(1)	検討委員会の委員	概要説明を行い、id/pwdを配布済み
(2)	実証実験実務者会議の委員	id/pwdを配布済み
(3)	運用検討会議の委員	id/pwdを配布済み
(4)	JEPA	11/17セミナーでご紹介／協力
(5)	JAGAT	11/22セミナーでご紹介／協力
(6)	日印産連	11/24セミナーでご紹介／協力
(7)	indexfont研究会	11/24セミナーでご紹介／協力
(8)	書協、雑協、電書協	11/30ご紹介／協力
(9)	電流協	12/13ご紹介／協力
(10)	IVS技術促進協議会	12/22セミナーでご紹介

5. 同意書送付先とお問い合わせ先

実証実験にご参加いただける方は、専用の同意書にご記入いただいて、PDFにて次のメールアドレスまでお送りください。

同意書送付先	gi-info@toppan.co.jp

- ツールのダウンロードサイト
- 字形共通基盤URL
- 字形共通基盤アクセスid/pwd

をご案内させていただきます。
また、お問い合わせに関しましても、上記メールアドレスまでお問い合わせをお願いいたします。

4.

Q&A