

Cahier des clauses techniques particulières

Prestations de numérisation de documents

Marché N° 251000006

Table des matières

1. Introduction.....	3
2. Description de la prestation attendue	3
2.1. Personnes ressources	3
2.2. Type de documents	3
2.3. Documentation à destination du titulaire.....	3
2.4. Enlèvement et restitution des lots.....	4
2.5. Numérisation des documents.....	5
2.6. Extraction ou saisie des métadonnées	6
3. Livrables.....	8
4. Critères de validation	8
5. Annexe : exemple de fichier XML à livrer.....	9

1. Introduction

L'Ifremer souhaite sous-traiter la numérisation et l'océrisation de documents (publications scientifiques, monographies, revues, littérature grise, manuscrits, tapuscrits) dans la perspective de leur publication dans son archive ouverte institutionnelle, Archimer (<http://archimer.ifremer.fr>). Le titulaire ne conserve aucun droit sur le produit de la numérisation : fichiers PDF, fichiers XML et métadonnées. L'Ifremer en dispose pour tout usage et en particulier la publication ouverte en base de données accessible à tout public et permettant le téléchargement des fichiers et des métadonnées.

2. Description de la prestation attendue

2.1. Personnes ressources

Au service Information Scientifique et Technique de l'Ifremer, les personnes ressources sont :

Nantes : Valérie Thomé et Marielle Bouldé, elles peuvent être contactées via l'alias bib.nantes@ifremer.fr

Brest : Doriane Ibarra qui peut être contactée par courriel à l'adresse doriane.ibarra@ifremer.fr.

Le titulaire doit fournir un interlocuteur unique dès l'établissement du devis et jusqu'à la réception de la commande. Aucune sous-traitance pour tout ou partie de la commande ne pourra être engagée par le titulaire sans l'accord exprès de l'Ifremer.

Si ce référent technique ne peut être le référent administratif, le titulaire devra également fournir le contact d'un référent administratif.

2.2. Type de documents

Il s'agira de numériser essentiellement de la littérature grise (rapports, etc.) ou encore des documents publiés par les éditions Ifremer, des revues anciennes publiées par l'ISTPM, des ouvrages anciens, des manuscrits ou tapuscrits.

2.3. Documentation à destination du titulaire

Un tableau de récolement (type tableur, csv,...) sera fourni par le Service Information Scientifique et Technique de l'Ifremer. Il détaillera le contenu des lots et le type de prestation

attendue. Il comportera une ligne descriptive par document. Les informations seront présentées en colonne, la première indiquera l'identifiant du document, cet identifiant sera reporté en page de garde de l'original¹. Les autres colonnes apporteront des éléments de type imprimé, tapuscrit vs manuscrit, format (A4, A3), recto/verso, recto, partiellement recto/verso, orientation des pages, langue(s) de rédaction, massicotage ou non, reliure ou non², nombre de pages... S'il y a des particularités, elles seront également mentionnées (feuillets volants ou hors format, calques, dépliants,...). L'éventuelle rareté des documents confiés sera également mentionnée (unica, manuscrits, reliures cuir...)

Ce tableau de récolement comportera une colonne Remarques pouvant être renseignée par le titulaire s'il rencontre une difficulté ou souhaite signaler une particularité.

Une autre colonne sera destinée au titulaire qui devra y reporter le nom du fichier PDF correspondant au document numérisé (une ligne du tableau correspondant à un document).

Il est attendu que le titulaire établisse une proposition de tarification sur la base du Bordereau des prix unitaires, dans la quinzaine suivant la réception de ce tableau. Après accord de l'Ifremer, un bon de commande est ensuite adressé au Titulaire dans les conditions définies dans le CCAP.

2.4. Enlèvement et restitution des lots

Les lots de documents à numériser seront à retirer par La Poste, un service de messagerie, un transporteur, un coursier...au Service Information Scientifique et technique de l'Ifremer :

- Direction Scientifique, Bâtiment Bougainville, Centre Ifremer Bretagne, 1625 Route de Sainte-Anne, 29280 Plouzané (horaires du lundi au vendredi : 8h-16h)– 29280 Plouzané (horaires : du lundi au jeudi : 9h 17 h ; vendredi 9h-16h)

ou

- Service Information Scientifique et Technique – Centre Ifremer Atlantique – rue de l'Île d'Yeu – 44300 Nantes (horaires : du lundi au vendredi : 9h-16h)

Des cartons³ vides et adaptés au transport de documents sont fournis par le prestataire au Service Information Scientifique et Technique en amont de l'enlèvement.

¹ Cela pourra être une cote, un code à barres, un numéro d'inventaire...

² Les documents seront triés selon ces critères à massicoter, à ne pas massicoter ; à relier, à rendre massicoté

³ Sous la désignation « carton », tout type de conditionnement sera accepté, il peut s'agir de caisses ou autre.

Les documents seront conditionnés en carton et préparés par le Service Information Scientifique et Technique. Le titulaire aura la charge du transport aller, du reconditionnement après numérisation et du retour.

Des emballages réutilisables, en matériaux recyclés et recyclables sont à privilégier.

2.5. Numérisation des documents

Les documents sont à traiter dans l'ordre des lignes du tableau de récolement. Ils doivent être reconditionnés en carton dans ce même ordre.

Avant d'être numérisés, sauf précision contraire, les documents sont à massicoter. En cas de massicotage certains documents devront être ré-encollés ou reliés selon mention dans le tableau de récolement. D'autres documents devront être restitués massicotés, auquel cas chacun de ces documents sera conditionné dans une chemise élastiquée ou une enveloppe individuelle.

Certains documents patrimoniaux ne pourront être massicotés. Ils supposeront une numérisation manuelle sur un scanner spécifique permettant de préserver l'intégrité de la reliure originale (angle d'ouverture à 120° maximum).

Il n'est pas attendu de fac-similé, sauf précision contraire dans le tableau de récolement.

Toutes les pages blanches non paginées doivent être supprimées, sauf précision contraire.

Selon indication dans le tableau de récolement, le format original des documents doit être respecté (A5, A4, A3,...) Les documents sont de format majoritairement A4.

Le cadrage est mono-page sauf précision contraire.

Les documents sont numérisés couvertures comprises, en couleur, en 300 dpi. Exceptionnellement, une autre définition pourra être attendue, le tableau de récolement le précisera.

Les pages doivent être redressées, au besoin, ou réorientées.

Les alignements doivent être redressés. La référence est la ligne de texte qui doit être horizontale, à la perpendiculaire des bords gauche-droit de la page.

Les bordures de page non nettes doivent être rognées.

Un fichier PDF/A 3a (cf <https://facile.cines.fr/>) doit être généré pour chaque document numérisé. Les fichiers sont nommés de manière incrémentale (le pas est de +1) en commençant par 001 pour le premier fichier. Ex :

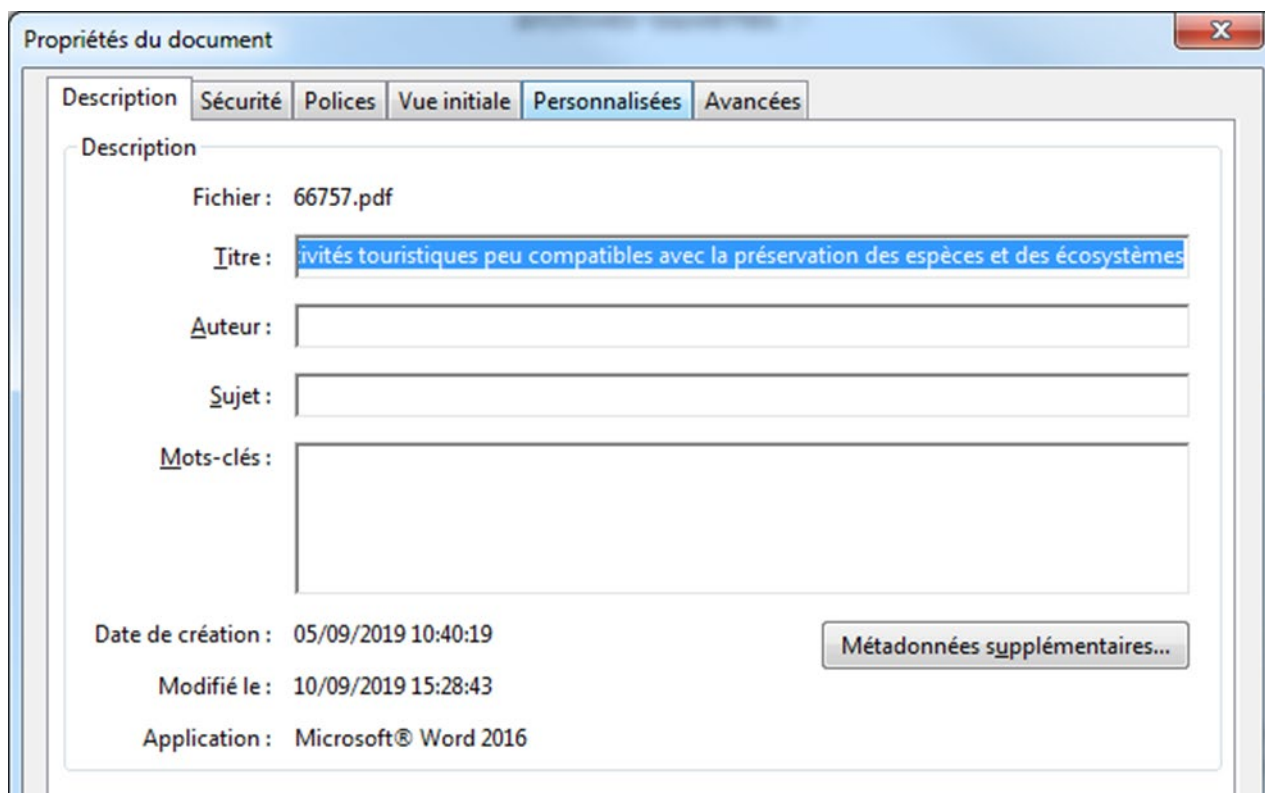
- 001.pdf
- 002.pdf
- ...

Un traitement de reconnaissance de caractère (OCR) est appliqué à l'intégralité de chaque PDF. Pour le texte intégral, il est attendu une qualité proche du standard de l'édition, soit

1/10000⁴ (1 erreur maximum sur 10 000 caractères). Le contrôle humain nécessaire à la qualité suppose une compétence linguistique (documents majoritairement en français et anglais). Il est attendu 100% d'exactitude dans les métadonnées (fichier XML), dans le texte des 1^{ère} et 4^{ème} de couverture et en page de garde du document numérisé (fichier PDF) ainsi que dans les propriétés du document PDF.

Le titre du document est enregistré dans les propriétés du document PDF.

Exemple :



2.6. Extraction ou saisie des métadonnées

L'ensemble des métadonnées listées ci-dessous est collecté dans le texte intégral et enregistré dans un fichier XML par lot de documents numérisés. Un exemple est fourni en annexe. Le format des champs est présenté ci-dessous⁵ :

⁴ Abdel Belaïd, Hubert Cecotti. Reconnaissance de caractères : évaluation des performances. In Mullot, Rémy. *Les documents écrits: de la numérisation à l'indexation par le contenu*, HERMES, 2006, Traité IC2, série informatique et systèmes d'information. [\(inria-00110927\)](#), p. 15

⁵ 1 : champ obligatoire mono occurrent

0..1 : champ optionnel mono occurrent

1..n : champ obligatoire multi occurrent

fichier (occurrence : 1) : nom du fichier PDF correspondant
annee (occurrence : 1) : année de publication au format YYYY
langue_texte_integral (occurrence : 1) : fra ou eng. Langue de rédaction du texte du document numérisé
titre_fr (occurrence : 0..1) : version française du titre du document
titre_en (occurrence : 0..1) : version anglaise du titre du document
resume_fr (occurrence : 0..1) : version française du résumé du document
resume_en (occurrence : 0..1) : version anglaise du résumé du document
auteur : (occurrence : 0..n) : auteur(s) du document
nom : (occurrence : 1) : nom de l'auteur
prénom : (occurrence : 1) : prénom de l'auteur

Les champs titre, résumé et auteurs sont contrôlés manuellement : si l'OCR contient des erreurs, ces erreurs sont toutes corrigées dans les métadonnées fournies (fichier XML). Ces données doivent également être exactes dans le texte intégral (1^{ère} et 4^{ème} de couverture, page de garde). Sur ces métadonnées, il est attendu 100 % d'exactitude.

Les titres, s'ils sont en majuscules dans le document PDF, sont transformés en minuscules dans le fichier XML à l'exception des sigles, de la première lettre du titre, de la première lettre des noms propres et des acronymes. Exemple :

ERIKA: SUIVI ECOTOXICOLOGIQUE DE LA QUALITE DES FLAQUES A TIGRIOPUS

➔ Erika: suivi écotoxicologique de la qualité des flaques à Tigriopus

Le fichier XML fourni est valide : il est possible de l'ouvrir et de le visualiser dans Google Chrome sans erreur. Dans cette perspective, les caractères interdits en XML sont encodés (ex : remplacement des < en <). Une attention particulière doit être apportée aux symboles (chimie, mathématiques...)

Le fichier PDF/A 3a fourni est valide : soumis au test Facile Cines (<https://facile.cines.fr/>), il est bien formé, valide et archivable dans PAC. Aucune tolérance d'erreur ne sera admise. Les fichiers doivent impérativement être corrigés avant livraison.

Les fichiers numérisés seront publiés sur Internet via Archimer pour être visualisés en ligne. La taille des fichiers PDF livrés doit donc être optimisée à l'aide d'une option équivalente à la fonctionnalité « Réduire la taille du fichier » du logiciel Acrobat. La taille des fichiers PDF ne devra pas excéder 10 Mo par fichier de 100 pages.

Un seul fichier XML contient l'ensemble des métadonnées d'un même lot de documents. Le lien entre PDF et métadonnées correspondantes dans le fichier XML est établi par le champ « fichier » du fichier XML où le nom du fichier PDF est renseigné, ex <fichier>001.pdf</fichier>.

0..n : champ optionnel multi occurrent

Les PDF et le fichier de métadonnées XML sont livrés à l'aide d'un service WEB de type filesender.

3. Livrables

Le livrable est composé des éléments suivants :

1. Les originaux reconditionnés en carton dans l'ordre défini par le tableau de récolement,
2. Chaque document massicoté doit être individuellement conditionné en chemise élastiquée ou enveloppe, ou ré-encollé,
3. Un unique fichier XML par lot de documents numérisés contenant les métadonnées les décrivant, livré au moyen d'un service WEB

Sur un serveur :

4. Un fichier PDF/A 3a par document numérisé,
5. Le tableau de récolement complété avec les noms des fichiers PDF reportés par le titulaire dans la colonne « nom de fichier » de la ligne correspondant au document ainsi que d'éventuelles remarques,
6. Une copie de sécurité doit être conservée par le titulaire tant que le Service Information Scientifique et Technique n'a pas réceptionné la livraison (délai de contrôle nécessaire cf § 4), cette copie pourra être exigée en cas de nécessité, elle sera détruite à la réception.

4. Critères de validation

Tous les documents originaux sont restitués intègres, selon les critères indiqués dans le tableau de récolement.

Un fichier PDF par document et un unique fichier XML par lot de documents (une même commande) sont livrés pour chaque commande.

Les fichiers PDF et XML correspondent aux attentes décrites dans le tableau de récolement et respectent les spécifications du présent cahier des charges cf § 32.5 et 2.6.

Les délais convenus en accord avec le titulaire et figurant dans le bon de commande sont respectés. Ce délai sera fixé en fonction du volume à traiter. Le délai maximum de réalisation entre commande et livraison est de 4 mois.

Le Service Information Scientifique et Technique effectuera un premier contrôle qualité portant sur le nombre de documents restitués, le nombre de fichiers, leur nommage, leur poids, leur validité...

Un second contrôle sera effectué par sondage : nombre, ordre, orientation, cadrage, netteté des pages, fidélité de l'océrisation, exactitude des métadonnées...

Chaque anomalie constatée sera signalée au titulaire au moyen d'une fiche afin qu'il puisse apporter les corrections nécessaires.

Le délai de signalement des anomalies est fonction du volume traité :

- 4 semaines à compter de la date de livraison pour un lot de moins de 50 documents ;
- 6 semaines à compter de la date de livraison pour un lot de 50 à 100 documents ;
- 8 semaines à compter de la date de livraison pour un lot de plus de 100 documents ;

Les corrections sont attendues dans les mêmes délais, leur livraison et validation conditionnent le règlement de la facture globale par l'Ifremer. Les coûts afférents aux corrections, transport compris, sont à la charge du titulaire et ne peuvent être facturés.

Les délais sont étendus d'une semaine en décembre et de quatre semaines en juillet et en août.

5. Annexe : exemple de fichier XML à livrer

```
<?XML version="1.0" encoding="UTF-8"?>
```

```
<liste>
```

```
  <document>
```

```
    <fichier>001.PDF</fichier>
```

```
    <annee>2005</annee>
```

```
    <langue_texte_integral>fra</langue_texte_integral>
```

```
    <titre_fr>Capacité d'enkystement, de survie et de germination et toxicité des principales espèces toxiques  
(affectant le littoral français) après transit stomacal chez Crassostrea gigas</titre_fr>
```

```
    <auteur>
```

```
      <nom>Laabir</nom>
```

```
      <prenom>Mohamed</prenom>
```

```
    </auteur>
```

```
    <auteur>
```

```
      <nom>Lassus</nom>
```

```
      <prenom>Patrick</prenom>
```

```
    </auteur>
```

```
  </document>
```

```
  <document>
```

```
    <fichier>002.PDF</fichier>
```

```
    <annee>2003</annee>
```

```
    <langue_texte_integral>fra</langue_texte_integral>
```

```
    <titre_fr>Etude des populations de tortues marines - bilan des actions scientifiques</titre_fr>
```

<auteur>

<nom>ROOS</nom>

<prenom>David</prenom>

</auteur>

</document>

<document>

<fichier>003.PDF</fichier>

<annee>2001</annee>

<langue_texte_integral>fra</langue_texte_integral>

<titre_fr>Restructuration du système de saisie, d'archivage et d'extraction des données du Réseau Mollusques des Rendements Aquacoles de l'IFREMER (REMORA)</titre_fr>

<resume_fr>Le réseau REMORA suit chaque année, depuis 1993, la survie, la croissance et la qualité de lots d'huîtres creuses sur le littoral français. Depuis cette date, la saisie, l'archivage et le traitement des données du réseau se font par le biais d'un ensemble de feuilles de calcul EXCEL 5 préformatées, associées à un programme d'aide créé sous Visual Basic pour Applications.

Devenu obsolète et trop rigide, le système a été modifié: la feuille de saisie de données a subi une refonte afin de coller au plus près aux souhaits des utilisateurs. L'archivage se fera grâce à une base de données ACCESS susceptible d'intégrer divers suivis autres que REMORA. Plus qu'une modification, il s'agit d'une étape qui s'inscrit dans la perspective d'une mise en réseau des données, et, plus tard, de l'intégration éventuelle à un système de gestion des données de plus grande ampleur.</resume_fr>

<resume_en>The REMORA Network studies year by year, since 1993, the survival, the growth and the quality of groups of copped oysters on French coasts. Since this year, the capture, the storage and the dataprocess are made in preformatted worksheets for EXCEL 5.0, with the help of a program created with Visual Basic for Applications.

Because of its stiffness, the process has been modified: the capture worksheet has been improved in order to fit the users' wishes. The data will be stored in an ACCESS database which will be able to include other studies than REMORA. This work is more than a modification: it's a step that will lead to a network base, and later, that may be included in a wider data management system.</resume_en>

<auteur>

<nom>LUCIANI</nom>

<prenom>Anthony</prenom>

</auteur>

</document>

<document>

<fichier>004.PDF</fichier>

<annee>2011</annee>

<langue_texte_integral>eng</langue_texte_integral>

<titre_en>Vulnerability of coastal ecosystems to global change and extreme events</titre_en>

</document>

<liste>