MODIFICATION 4

QUESTIONS ET RÉPONSES

Q15. Le diagramme de l'annexe 1 représente-t-il le flux de travail *actuel* ou le résultat *attendu*? Le lien entre l'élément « Fichier de validation » et l'OCR/ICR vise-t-il à illustrer quelque chose et que sont les trois bases de données liées à l'élément « Exportation image et données »?

R15. L'annexe 1 représente le flux de travail de l'imagerie *actuel* et indique des endroits possibles pour le processus de rectification de l'accessibilité. La validation des données peut être réalisée pour appuyer l'OCR/ICR et tout type de numérisation de formulaire ou de code à barres. Les trois bases de données liées à l'élément « Exportation image et données » sont des dépôts utilisés pour montrer des possibilités, comme le stockage d'images brutes pour les originaux de preuves documentaires ou le stockage d'images sur un serveur sécurisé accessible pour les clients.

Q16. Pouvez-vous donner des exemples de documents et de résultats associés au flux de travail actuel présenté à l'annexe 1 pour le flux de travail et les technologies actuels?

R16. Les résultats obtenus en ce moment sont souvent des documents PDF ou PDF/A (selon les besoins des clients) ainsi que des documents XML connexes comprenant diverses métadonnées (y compris la numérisation de renseignements aux fins de preuves documentaires et tout renseignement extrait de documents numérisés). Des exemples pourraient être fournis à l'étape du projet pilote.

Q17. Pouvez-vous expliquer davantage le but de l'OCR et de l'ICR ainsi que leur lien avec l'élément « Fichier de validation » présenté à l'annexe 1 pour le flux de travail et les technologies actuels?

R17. La validation des données consiste habituellement à accéder à une base de données ou à un fichier de référence pour déchiffrer des codes à barres, valider des données en fonction des numéros de dossier ou valider la précision de l'information extraite.

Q18. Pouvez-vous expliquer davantage le but de l'OCR et de l'ICR ainsi que leur lien avec l'élément « Validation auprès d'un système client » présenté à l'annexe 1 pour le flux de travail et les technologies actuels?

R18. Veuillez consulter les réponses fournies aux questions 15 et 17.

Q19. Pouvez-vous préciser quels articles de la norme de l'Union européenne EN 301-549 (2018) doivent être respectés pour la solution?

R19. Les articles particuliers propres à ce défi sont les articles 9 (Web) et 10 (Non-Web documents).

Q20. Comment vérifiez-vous actuellement si un document HTML5 ou ePub3 est conforme?

R20. Un outil de validation devrait être fourni dans le cadre du processus d'assurance de la qualité.

Q21. Qu'utilisez-vous comme interface de rectification et outils de révision dans le cadre du flux de travail actuel?

R21. Nous ne produisons actuellement aucune interface de rectification de l'accessibilité et il est donc essentiel d'intégrer cet outil à la solution proposée.

Q22. Le schéma de l'annexe 1 qui se trouve dans la DP est-il essentiellement un diagramme conceptuel qui montre comment le CSID voit la conception de la solution ou est-ce qu'il s'agit de la mise en œuvre exigée?

Quels processus présentés dans le schéma sont déjà en place?

R22. Tous les processus indiqués sont actuellement en place et représentatifs de notre flux de travail opérationnel en cours, à l'exception des deux encadrés bleus (« Utilisation de l'IA » et « Conversion à un format accessible ») indiquant les possibilités de mise en œuvre de l'accessibilité. Ils ne doivent pas être considérés comme les seules options pour la mise en œuvre; nous sommes ouverts à toutes les possibilités respectant nos exigences.

Q23. À l'étape de l'importation, pouvez-vous décrire le courrier entrant, les formulaires, les documents d'archives, les documents reçus par fax et les documents électroniques faisant partie des entrants du système actuel?

R23. Les documents à numériser sont très variés. Il peut s'agir de dossiers propres et bien organisés ainsi que de boîtes de lettres ou de documents en format légal ou de documents endommagés, reliés ou agrafés, et non structurés. Ces documents comprennent aussi le courrier (notamment des pièces de correspondance, des formulaires, des chèques, etc.) et nécessitent souvent des mesures initiales de classement et de classification avant d'être numérisés.

Q24. Pour l'étape de l'importation, pouvez-vous donner des descriptions individuelles et des exemples représentatifs pour chaque type de documents? Ça serait très utile à notre analyse.

R24. Publications : publications reliées ou non. Les publications reliées peuvent être numérisées à l'aide de numériseurs de livres ou de numériseurs de table, ou être détachées et numérisées au moyen de numériseurs à haute vitesse d'IntelliScan.

Formulaires gouvernementaux : nombreux modèles de formulaires provenant d'employés ou de citoyens. Les formulaires peuvent comprendre des zones de texte ou d'autres renseignements qui doivent être extraits et insérés dans des documents XML.

Correspondance du programme : le courrier entrant comprend la correspondance qui est habituellement numérisée et traitée grâce à l'OCR/ICR.

Collections archivées : la majeure partie de notre volume provient de collections archivées comprenant habituellement un grand nombre de documents similaires (fichiers RH, documents d'enquête de sécurité, dossiers archivés, données météorologiques ou scientifiques, etc.). L'extraction d'information dépend des besoins du client et de l'utilisation des documents numérisés. Les collections peuvent être numérisées pour être archivées ou intégrées à une base de données.

Fax : il peut s'agir de formulaires, de correspondance et de communications opérationnelles générales.

Q25. À l'étape de l'importation, veuillez indiquer s'il y a des concordances entre les documents entrants de sorte qu'ils puissent être automatiquement considérés comme appartenant à un type de documents particulier aux fins de traitement en aval. Par exemple, une part importante du courrier entrant peut être constituée d'un certain type de formulaires remplis qui, par conséquent, peuvent être reconnus et traités d'une façon particulière. Est-ce l'objectif de l'étape de la classification automatique? Le cas échéant, y a-t-il des configurations de traitement propres à l'OCR ou à l'ICR pour chaque type de document déterminé?

R25. En général, plusieurs processus d'imagerie se déroulent simultanément tout au long de notre procédure. Chaque collection possède un certain degré d'uniformité, qui peut être très spécifique (un seul formulaire reçu par télécopieur et dont nous pouvons extraire la même information à répétition) ou très variable (p. ex. archives des RH comprenant différents documents regroupés dans des classeurs et des boîtes, pour lesquels l'OCR n'offre qu'une fonction de recherche, mais sans extraction de données). Les projets sont généralement identifiés tout au long du processus au moyen d'en-têtes et de séparateurs, lesquels indiquent au système d'imagerie (numériseurs et systèmes de soutien) quels profils doivent être utilisés. L'identification initiale des propriétés et de l'origine du document est généralement réalisée manuellement, en fonction de son origine (case postale pour le courrier entrant, identification de la boîte de dossiers ou

Q26. À l'étape de l'importation, veuillez fournir autant d'exemples de types de documents pertinents que possible ainsi que le pourcentage approximatif du volume total qu'ils représentent.

R26. Des exemples seront fournis pour appuyer l'élaboration du projet pilote à la phase 1.

Q27. Existe-t-il des concordances d'images dans ou entre les types de documents qui pourraient permettre d'automatiser l'ajout de texte de remplacement dans certaines situations? Le cas échéant, quel pourcentage du texte de remplacement de l'image pourrait être reconnu et appliqué automatiquement en ce moment? Quel est votre objectif relativement à l'application de texte de remplacement assistée par IA pour ce type de documents dans l'avenir?

R27. La cohérence entre les documents se trouve seulement à l'intérieur de collections particulières et pourrait ne jamais être applicable à d'autres collections à venir. En outre, comme nous travaillons plus précisément avec des publications gouvernementales, la plupart des images que nous traitons sont des graphiques, des diagrammes et des tableaux. Nous cherchons donc des solutions qui utilisent l'IA de façon à décoder adéquatement ces images riches en renseignements et à fournir des données dans un format accessible.

Q28. Combien de pages sont traitées en moyenne par année et par mois en ce moment et combien de pages devraient être traitées dans l'avenir? Y a-t-il des périodes de pointe dont nous devrions être informés?

R28. Nous traitons jusqu'à 40 millions d'images par année. Notre volume dépend des collections majeures numérisées ainsi que des fluctuations du courrier entrant de nos clients. Il est difficile d'évaluer le volume de documents qui devront faire l'objet d'un processus de rectification de l'accessibilité en ce moment, mais nous nous attendons à recevoir de plus petites collections et des documents qui ont une certaine valeur opérationnelle plutôt que de vastes collections d'archives.

Q29. En moyenne, combien d'heures-personnes de travail sont requises par document en ce moment après la numérisation? Quel est votre objectif au chapitre des futures heures-personnes requises par document lorsque ce projet sera terminé?

R29. Actuellement, selon le degré d'assurance de la qualité et de saisie de données exigé par le client, le traitement prend de cinq secondes à cinq minutes par page. Nous n'avons pas d'objectif fixé en ce qui a trait à la rectification de l'accessibilité, mais étant donné notre modèle d'imagerie à fort volume, nous avons l'intention de réduire au minimum l'intervention humaine de façon à diminuer le coût par page et à accélérer le processus de rectification.

Q30. Pouvez-vous préciser quels types de documents (images) feront l'objet de ces traitements?

- **Des documents de gestion ou des communications récentes : factures, états, certificats, contrats, courrier, etc.?**
- **Des formulaires remplis à la main et dactylographiés?**
- **Des documents patrimoniaux : lettres, registres, plans, livres, journaux, etc.?**

R30. Veuillez consulter les réponses fournies aux questions 23, 25 et 31.

Q31. Pouvez-vous préciser de quels types de supports physiques sont tirées les images et comment celles-ci ont été numérisées (matériel de numérisation, résolution, profondeur : niveau de gris, couleurs)?

R31. Les documents à numériser sont très variés. Il peut s'agir de dossiers propres et bien organisés ainsi que de boîtes de lettres ou de documents en format légal ou de documents endommagés, reliés ou agrafés, et non structurés. Ces documents comprennent aussi le courrier (notamment des pièces de correspondance, des formulaires, des chèques, etc.) et nécessitent souvent des mesures initiales de classement et de classification avant d'être numérisés. Les caractéristiques de la numérisation peuvent varier selon les besoins du client et le type de document à numériser, mais ce sont habituellement des images bitonales ou à échelle de gris 8 bits pour les documents dactylographiés, et des images couleur 24 bits pour les documents qui comprennent des images, des graphiques ou des cartes, et la résolution varie de 200 à 300 ppp.

Q32. Avez-vous des exemples illustrant ces types de documents?

R32. Des exemples seront fournis pour appuyer l'élaboration du projet pilote à la phase 1.