



Comment la GenAI améliore l'extraction d'informations de documents industriels complexes

29 juin 2026, 12:00

Sreeraj Rajendran

Nicolás González-Deleito

Ali Beigrezaei

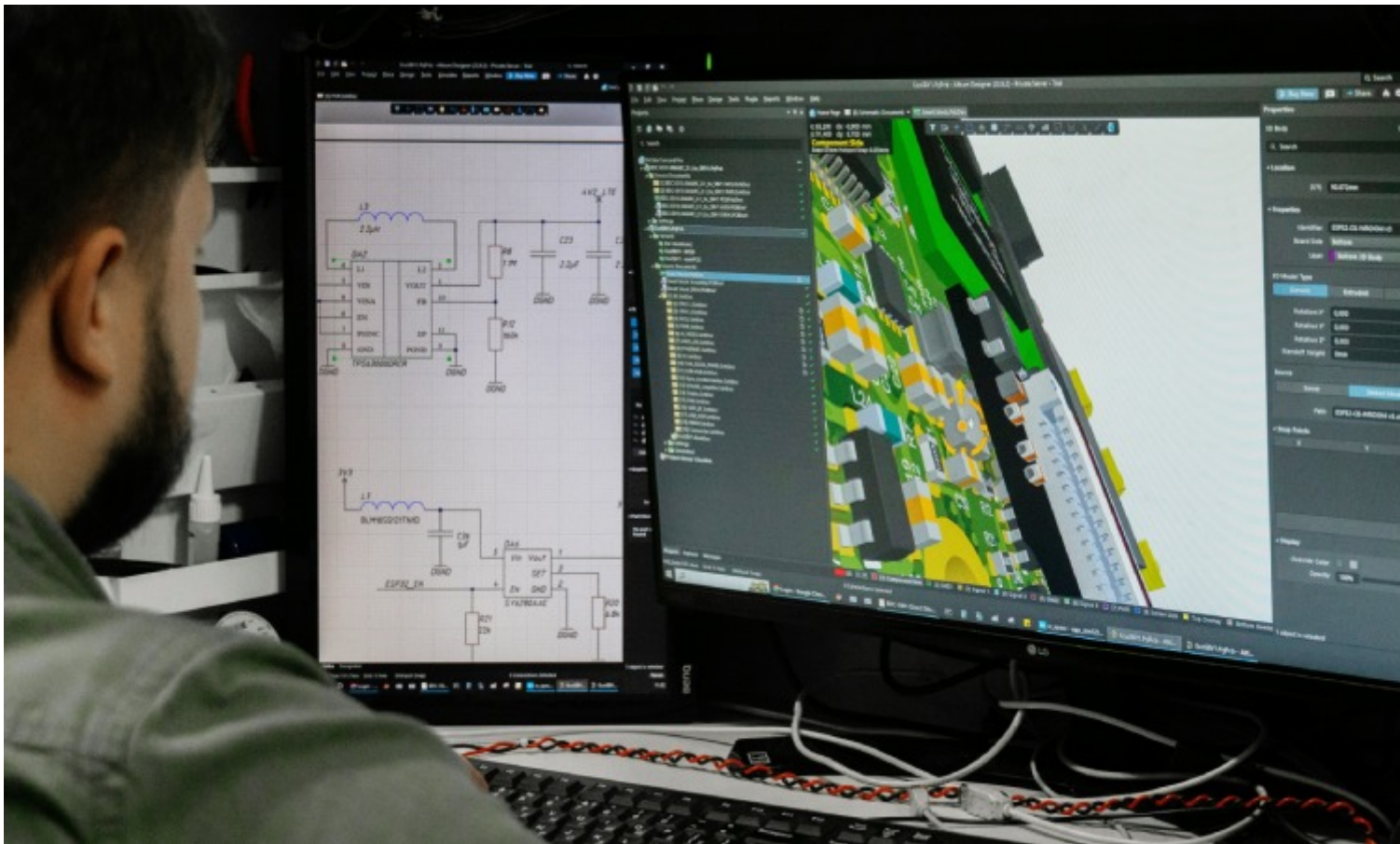
Aziz Nebli

Des documents techniques aux informations exploitables

Les processus industriels, l'implantation des usines, la configuration des sites de production ou les spécifications des composants sont généralement documentés dans un large éventail de sources techniques?: descriptions textuelles, schémas, figures, tableaux et images. Extraire les informations pertinentes de ces documents se révèle essentiel pour des activités telles que la maintenance, les audits, les certifications et la création de jumeaux numériques. Cependant, identifier, relier et interpréter de manière fiable les informations provenant de ces sources hétérogènes reste une tâche complexe.

Dans le cadre du projet international [NARRATE](#), mené avec le soutien de la Région de Bruxelles-Capitale et Innoviris, Sirris et 3E étudient comment l'IA générative peut extraire, relier et structurer automatiquement ces informations, ouvrant ainsi la voie à des flux de traitement et

d'automatisation plus rapides et plus efficaces.



À propos de NARRATE

[NARRATE](#) est un projet de recherche international rassemblant quatorze partenaires issus de cinq pays : Belgique, Corée du Sud, Espagne, Estonie et Turquie. Ce projet vise à développer des interfaces d'interaction homme-machine fiables et éthiques reposant sur l'IA, en combinant l'intégration de données multimodales et la mise à jour automatisée des connaissances.

[En savoir plus](#)

Pourquoi l'extraction d'informations reste-t-elle difficile??

La complexité ne réside pas dans la recherche d'informations isolées, mais dans la compréhension de la manière dont elles sont liées à travers différentes sources?:

- Souvent, les données ne sont pas structurées et sont présentées dans différents formats (texte, tableaux ou images, par exemple)
- Des informations dans différents formats sont liées entre elles (une description textuelle fait référence à des informations contenues dans un tableau ou une image, par exemple)
- Les références sont très éloignées les unes des autres au sein d'un même document ou réparties dans plusieurs fichiers
- Un même concept peut être décrit à l'aide de différentes conventions de nommage ou de différents identifiants

L'extraction d'informations ne constitue donc qu'une partie du problème. Les systèmes doivent aussi identifier les relations, regrouper les informations et préserver le contexte au sein de groupes de documents hétérogènes.

Transformer des documents hétérogènes en données exploitables

Dans le cadre du projet [NARRATE](#), Sirris et 3E développent des pipelines basés sur l'IA capables d'extraire, de relier et de structurer automatiquement des informations provenant d'un large éventail de documents d'ingénierie. Pour y parvenir, Sirris et 3E associent deux stratégies d'IA complémentaires.

1. LLM locaux

Sirris se concentre sur les grands modèles de langage (LLM) hébergés localement, tels que LLaMA, Gemma et Mistral. Conserver les modèles sur l'infrastructure interne garantit que les données clients sensibles ne quittent pas l'environnement de l'entreprise. Le flux de travail initial suit une approche en plusieurs étapes.

- 1. Identifier le type et le contenu du document.
- 2. Extraire les informations pertinentes.
- 3. Structurer les données extraites.
- 4. Répéter les étapes ci-dessus pour chaque document.
- 5. Regrouper les informations provenant de différents fichiers dans un seul fichier de sortie.

Au fur et à mesure de l'avancement du projet, il s'est avéré particulièrement difficile de relier les informations issues de plusieurs documents. L'approche s'est alors déplacée vers un flux de travail basé sur des agents, dans lequel l'IA coordonne les différentes étapes nécessaires à la détection, la mise en relation et la validation des informations au sein d'un ensemble complet de documents.

Ce flux de travail agentique permet au système de traiter des ensembles de documents polyvalents et multimodaux, comprenant à la fois du texte, des tableaux, des figures et des schémas, dans des formats et selon des conventions de nommage variés.

2. LLM basés dans le cloud

Parallèlement, 3E développe un pipeline basé dans le cloud au sein d'un environnement Azure sécurisé. Les documents sont d'abord traités et segmentés de manière intelligente à l'aide de

technologies telles qu'Azure Computer Vision et Azure Document Intelligence. Des modèles de GenAI identifient ensuite les types d'informations et tentent de reconstituer automatiquement les liens qui les unissent. L'objectif est le même?: transformer une documentation technique complexe en une vue consolidée structurée, en limitant autant que possible les interventions manuelles.



Améliorer la fiabilité grâce à la validation

Afin d'améliorer la fiabilité, les deux partenaires ont développé des mécanismes de validation approfondis. Sirris utilise Langfuse pour surveiller les invites (« prompts » en anglais), comparer les paramètres du modèle et évaluer la qualité des résultats. 3E utilise une approche d'ensemble dans laquelle plusieurs LLM effectuent en parallèle la même tâche d'extraction. Les résultats obtenus font ensuite l'objet d'une vérification croisée visant à détecter les incohérences, les informations manquantes et les doublons avant la consolidation finale.

Comment mesurer le succès??

Une question importante restait en suspens?: comment évaluer la qualité des informations extraites automatiquement?? Les deux partenaires ont développé des métriques d'évaluation sur mesure axées sur les éléments suivants?:

- Exhaustivité
- Présence des éléments requis
- Qualité sur le terrain

- Exactitude numérique

Les premiers résultats sont prometteurs. L'extraction d'informations donne déjà d'excellents résultats. Cependant, traiter automatiquement les relations complexes entre des informations reste la partie la plus difficile.

Quelles sont les prochaines étapes??

La première phase du projet a déjà démontré que la GenAI peut significativement accélérer l'extraction d'informations techniques de documents d'ingénierie complexes. La prochaine étape consiste à étendre encore ces capacités en permettant la liaison d'éléments complexes et en traduisant les connaissances extraites en une aide concrète à la décision. Le consortium explorera également des applications telles que les tableaux de bord intelligents, les systèmes de recommandation et la gestion automatisée des connaissances, destinées à faciliter la prise de décision opérationnelle.

Le parcours est loin d'être terminé. Toutefois, les progrès réalisés dans le cadre du projet [NARRATE](#) montrent déjà comment une documentation technique complexe peut être transformée en informations numériques exploitables. À mesure que la GenAI continue de mûrir, l'extraction d'informations fiables pourrait devenir un facteur clé pour améliorer les processus, notamment en accélérant l'intégration, en renforçant l'évolutivité et en optimisant la gestion des actifs.

Vous vous demandez comment la GenAI pourrait vous aider à rationaliser vos propres flux d'ingénierie??

Sirris aide les entreprises à mettre en œuvre l'IA et la GenAI afin de relever des défis industriels complexes allant de l'extraction d'informations et la gestion des connaissances à l'automatisation et l'aide à la décision.

[Contactez-nous](#)

Partners

avec le soutien d'ITEA4, d'Innoviris et de la Région de Bruxelles-Capitale



RÉGION DE BRUXELLES-CAPITALE
BRUSSELS HOOFDSTEDELIJK GEWEST
BRUSSELS-CAPITAL REGION

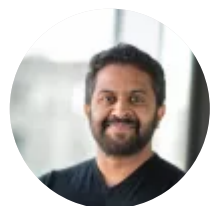


Cofinancé par



RÉGION DE BRUXELLES-CAPITALE
BRUSSELS HOOFDSTEDELIJK GEWEST
BRUSSELS-CAPITAL REGION

Auteurs



Sreeraj Rajendran



Nicolás González-Deleito



Ali Beigrezaei



Aziz Nebli