Le langage SGML: aux origines des langages de balisage

Stéphane FOSSE
fosse.fr

21 novembre 2025

En octobre 1986, l'Organisation internationale de normalisation publiait la norme ISO 8879, consacrant officiellement le Standard Generalized Markup Language. Trente ans avant que les navigateurs modernes ne deviennent des outils du quotidien, ce langage de balisage posait les fondations techniques de ce qui deviendrait HTML et XML. L'histoire de SGML est celle d'une révolution silencieuse dans la manière dont les machines traitent et comprennent les documents.

Tout commence dans les années 1960, au cœur des laboratoires d'IBM. Charles Goldfarb, juriste de formation passé à l'informatique, travaille alors au Cambridge Scientific Center du Massachusetts. Il collabore avec deux collègues, Edward Mosher et Raymond Lorie. En 1969, les trois hommes développent ce qu'ils baptisent GML, pour Generalized Markup Language. Le sigle n'est pas un hasard : il correspond aux initiales de leurs trois noms de famille. Ce clin d'œil onomastique restera gravé dans l'histoire de l'informatique.

Le problème que Goldfarb cherchait à résoudre relevait d'une frustration concrète. Les systèmes de traitement de texte de l'époque mélangeaient allègrement le contenu des documents avec les instructions de mise en forme. Chaque système utilisait ses propres codes propriétaires, rendant les documents prisonniers de leurs logiciels d'origine. Un texte créé sur une machine devenait illisible sur une autre. Cette situation paralysait les grandes organisations qui devaient gérer des masses documentaires considérables et les conserver sur de longues durées.

L'intuition de Goldfarb tenait en deux principes qui, formulés noir sur blanc dans l'annexe A.1 de la future norme ISO, allaient transformer la gestion documentaire. Le balisage devait être déclaratif et non procédural : au lieu d'indiquer à la machine comment formater un passage, on lui indiquerait ce que ce passage représentait. Un titre resterait un titre, qu'on l'imprime en gras, en italique ou en lettres capitales. Le second principe exigeait la rigueur : les structures documentaires devaient être définies avec la même précision que les structures de données dans les programmes informatiques.

GML prenait la forme d'un ensemble de macros pour SCRIPT, le formateur de texte d'IBM intégré au Document Composition Facility. Les balises ressemblaient déjà à ce que nous connaissons, avec leurs chevrons caractéristiques. Un chapitre s'ouvrait par une balise, un paragraphe par une autre. La structure logique du document devenait explicite, séparée de sa présentation visuelle.

En 1974, Goldfarb entreprit de généraliser ces concepts au-delà du contexte IBM. Il esquissa les contours de ce qui deviendrait SGML. L'objectif dépassait largement le cadre d'un produit commercial : il s'agissait de créer un métalangage capable de définir n'importe quelle famille de documents structurés. Un langage pour créer des langages.

De la recherche à la normalisation internationale

Le chemin vers la standardisation fut long. Un premier projet de norme vit le jour en 1980, sous l'égide du comité ISO/TC 97 chargé des systèmes de traitement de l'information. Les discussions se poursuivirent pendant six années. En 1983, la Graphic Communications Association adopta la sixième ébauche comme standard industriel sous la référence GCA 101-1983. Des acteurs de poids commençaient à s'y intéresser : l'Internal Revenue Service américain, le département de la Défense des États-Unis. Ces adoptions précoces validaient l'approche mais révélaient aussi la nécessité d'une norme internationale reconnue.

En 1984, le groupe de travail fut reconstitué sous l'égide de l'ISO et de la CEI. L'année suivante paraissait un projet de norme internationale. Le 15 octobre 1986, la norme ISO 8879 était officiellement publiée, un document de 155 pages qui définissait dans ses moindres détails le Standard Generalized Markup Language. Un amendement suivit en 1988, puis deux correctifs en 1996 et 1999.

La norme SGML introduisait le concept de DTD, la Document Type Definition. Cette grammaire formelle décrivait la structure autorisée pour une famille de documents : quels éléments pouvaient apparaître, dans quel ordre, avec quels attributs. Un document conforme à une DTD donnée pouvait être vérifié automatiquement par

un programme, le *parseur*, qui s'assurait que toutes les règles étaient respectées. Cette validation systématique inaugurait une ère nouvelle dans la gestion documentaire.

Le département américain de la Défense devint l'un des plus fervents promoteurs de SGML. Dès février 1987, il lança le programme CALS, d'abord acronyme de Computer-aided Acquisition and Logistic Support, rebaptisé ensuite Continuous Acquisition and Life-cycle Support. En février 1988 paraissait le standard MIL-M-28001, imposant l'utilisation de SGML pour l'échange de documents techniques militaires. Ce standard, modifié pour la dernière fois en 1993, reste en vigueur. La norme MIL-STD-38784, consacrée aux manuels techniques, publiée en 1995 et mise à jour en 2018, repose toujours sur une DTD SGML.

L'industrie aéronautique, les secteurs ferroviaire et des télécommunications emboîtèrent le pas. L'Association of American Publishers développa sa propre application SGML pour l'échange de manuscrits entre auteurs et éditeurs, approuvée comme standard ANSI/NISO Z39.59 en décembre 1988, puis portée au niveau international comme ISO 12083 en 1994. La Text Encoding Initiative, consortium universitaire fondé en 1987, élabora des recommandations pour l'encodage des textes littéraires et linguistiques sous forme de DTD SGML.

DocBook naquit vers 1991 pour la documentation des logiciels Unix. Le monde de l'édition technique s'en empara. Robin Cover, chercheur et documentaliste, assembla une collection de ressources sur SGML qui devint une référence incontournable pour les praticiens de la discipline. Hébergées par OASIS, ces Cover Pages listaient des dizaines d'applications dans les domaines les plus variés.

Goldfarb lui-même publia en 1990 « The SGML Handbook » chez Oxford University Press, un ouvrage de 663 pages reprenant le texte intégral de la norme ISO avec des annotations extensives. Ce livre devint la bible des développeurs SGML, l'ouvrage vers lequel on se tournait quand les subtilités de la spécification résistaient à l'interprétation.

L'Oxford English Dictionary dans sa deuxième édition, parue en 1989, fut entièrement balisé avec un langage dérivé de SGML à l'aide de l'éditeur LEXX. Des millions d'entrées lexicographiques structurées selon une DTD spécifique, un monument éditorial converti en données manipulables par des programmes.

La puissance de SGML avait une contrepartie : sa complexité. La norme autorisait des syntaxes concrètes variées, la modification des délimiteurs, l'omission de balises dans certaines conditions, des minimisations diverses. Un parseur SGML devait interpréter une déclaration SGML définissant les caractères autorisés, les longueurs de noms, les capacités du système. Cette flexibilité, pensée pour s'adapter aux contraintes de systèmes hétérogènes, compliquait la mise en œuvre.

Le département de la Défense avait investi massivement dans des solutions SGML. Les éditeurs logiciels répondirent avec des produits onéreux, ciblant les grands comptes gouvernementaux et industriels. Un cercle vertueux pour certains, une barrière à l'entrée pour les autres. Les petites structures ne pouvaient s'offrir ces outils coûteux.

Et puis survint le Web. En décembre 1990, au CERN à Genève, Tim Berners-Lee et Robert Caillau présentèrent leur projet d'hypertexte global. HTML, le langage de balisage qu'ils avaient conçu pour structurer les pages web, s'inspirait directement de SGML. Berners-Lee avait explicitement voulu que HTML soit une application de SGML, même si les navigateurs de l'époque ne validaient pas les documents contre une DTD formelle. La simplicité l'emportait sur la rigueur.

HTML démontrait à grande échelle ce que les promoteurs de SGML avançaient depuis des années : le balisage sémantique fonctionnait, il avait une valeur économique et intellectuelle. Des millions de pages web prouvaient l'intérêt de séparer structure et présentation. Les balises `<h1>`, `<p>`, `<a>` que manipulent quotidiennement les développeurs web descendent en ligne directe des concepts élaborés par Goldfarb et ses collègues dans les années 1960 et 1970.

La version 2.0 de HTML, en 1995, se rapprocha davantage de SGML. La version 4, en 1997 puis 1999, se conformait pleinement à la norme ISO 8879, avec ses trois DTD pour les variantes Strict, Transitional et Frameset. Dans la pratique, cependant, les navigateurs ne traitaient pas HTML comme du SGML : ils avaient développé leurs propres règles de parsing, plus tolérantes aux erreurs, plus adaptées à la réalité des pages web mal formées qui pullulaient sur Internet.

En parallèle, le consortium W3C reconnaissait que SGML dans sa version complète ne convenait pas au Web. Les navigateurs ne le prendraient jamais en charge nativement. Trop de fonctionnalités optionnelles, trop de variantes possibles. En 1996 débutèrent les travaux sur une version simplifiée, un profil de SGML débarrassé de ses options les plus exotiques. Le terme « SGML-Lite » circulait informellement. XML, pour Extensible Markup Language, fut le nom retenu.

James Clark, développeur britannique qui avait créé SP, l'un des parseurs SGML les plus fiables et les plus utilisés, joua un rôle déterminant dans la conception de XML. En décembre 1997, il publia une note technique du W3C détaillant les différences entre SGML et XML. La nouvelle spécification désactivait nombre de fonctionnalités SGML : pas d'omission de balises, pas de balises abrégées, pas de sous-documents imbriqués, sensibilité à la casse des noms d'éléments. XML interdisait aussi les références à des entités externes dans les valeurs d'attributs, imposait la fermeture de toutes les balises, exigeait la présence d'un identifiant système dans les déclarations d'entités externes.

En février 1998, XML devenait une recommandation du W3C. La spécification précisait que tout document XML devait être un document SGML conforme. L'annexe C du standard XML établissait la correspondance formelle. XML n'était pas un remplacement de SGML, mais un sous-ensemble rigoureusement défini.

La migration s'accéléra. La Text Encoding Initiative proposa une version P4 de ses recommandations avec un choix entre SGML et XML, puis une version P5 exclusivement XML. DocBook suivit le même chemin : les versions 4.x supportaient les deux syntaxes, la version 5.0 abandonna SGML. Les outils se multipliaient côté XML tandis que le marché SGML se contractait.

Betty Harvey, consultante spécialisée dans les technologies documentaires, présentait en 2016 à la conférence Balisage un état des lieux significatif. Trente ans après la publication de la norme ISO, SGML restait utilisé dans certains contextes, principalement militaires. Le système JCALS du département de la Défense, développé dans les années 1990, fonctionnait toujours avec ses composants Arbortext Editor et Datalogics DL Composer. Des contrats exigeaient la livraison de données en SGML. Les organisations concernées devaient jongler entre des chaînes de production XML modernes et des exigences de sortie SGML héritées d'une époque révolue.

Quelques éditeurs logiciels maintenaient encore un support SGML : Justsystem avec XMetal, Adobe avec FrameMaker, PTC avec Arbortext Editor. Leur documentation marketing mentionnait à peine cette capacité, reléguée au rang de fonctionnalité de niche. Le parser SP de James Clark, devenu OpenJade, restait disponible dans les distributions Linux, mais son développement actif avait cessé depuis longtemps.

La Library of Congress, dans ses recommandations pour la préservation des formats numériques, continuait d'accepter SGML comme format valide pour les documents textuels, à condition qu'une DTD accompagne le document. Cette institution avait elle-même utilisé SGML dans les années 1990 pour le projet American Memory, avec une DTD maison pour la transcription de livres et documents historiques tombés dans le domaine public. Elle avait depuis converti ces fichiers en XML.

L'héritage de SGML dépasse largement son usage direct. HTML, même dans sa version 5 qui définit ses propres règles de parsing indépendantes de SGML, conserve la syntaxe des chevrons, des attributs entre guillemets, des entités nommées. XML structure des formats aussi variés que les documents Office Open XML de Microsoft, les fichiers SVG, les flux RSS et Atom, les messages SOAP. XSLT et XSL-FO, langages de transformation et de mise en forme, descendent de DSSSL, le Document Style Semantics and Specification Language standardisé en 1996 pour accompagner SGML. JSON lui-même, bien que syntaxiquement différent, répond au même besoin de structuration des données que SGML cherchait à satisfaire.

Charles Goldfarb, reconnu comme le « père de SGML » et le « grand-père de HTML et du World Wide Web » par l'IT History Society, a reçu le prix Gutenberg des Printing Industries of America et le titre de Fellow honoraire de la Society for Technical Communication. Son apport conceptuel irrigue l'ensemble de l'écosystème numérique contemporain, même si le nom SGML évoque désormais un passé révolu pour la plupart des développeurs.

En 2016, la conférence Balisage organisait une fête pour le trentième anniversaire de la norme. L'invitation résumait la situation avec une lucidité teintée d'ironie : « Notre écosystème informationnel actuel ignore largement SGML, voire totalement. Pourtant, il dépend entièrement de sa descendance : HTML et XML. SGML n'a pas fait grand bruit, mais les ondes qu'il a créées continuent de se propager. »

Références

[1] James Clark. Comparison of SGML and XML. W3C Note. World Wide Web Consortium, déc. 1997.

[2] Betty Harvey. SGML in the Age of XML. In : *Balisage: The Markup Conference*. T. 17. 2016.

[3] History of Information. IBM Introduces the Generalized Markup Language (GML). History of Information.

[4] History of Information. The SGML Standard is Accepted by the ISO. History of Information.

[5] ISO. Information processing – Text and office systems – Standard Generalized Markup Language (SGML). International Organization for Standardization, 1986.

[6] IT History Society. Dr. Charles F. Goldfarb. IT History Society.

[7] Library of Congress. Sustainability of Digital Formats: Standard Generalized Markup Language (SGML). Library of Congress.

[8] W3C. Overview of SGML Resources. World Wide Web Consortium.