خمسة عشر عامًا من الويب الدلالي: من التقنيات إلى الممارسات الجيدة

خوان أنطونيو باستور سانشيث



الملخص:

يناقش المقال نضج الويب الدلالي بعد سنوات من تطوير تقنياته، مؤكدًا أن نجاح نشر البيانات لا يعتمد على المعايير والأدوات وحدها. ويعرض أهمية البيانات الوصفية والمنشأ والجودة والتراخيص والإصدارات والمعرّفات المستقرة وآليات الوصول، بوصفها عناصر في دورة حياة المجموعة. ويتناول عمل فريق (W3C) للممارسات الجيدة، وربط النشر بحاجات المستهلكين والتغذية الراجعة والإثراء وإعادة الاستخدام. كما يشرح فوائد الفهم والمعالجة والاكتشاف والثقة والربط والتشغيل البيني، وضرورة التوازن مع الخصوصية والأمن. ويؤكد أن الإرشادات تحتاج مراجعة مستمرة مع تغير التطبيقات والتقنيات، وأن المعرفة التقنية يجب أن تصحبها رؤية لديناميات البيانات وإدارتها. وتبقى أعداد الممارسات وحالة الوثائق والمفردات مرتبطة بمرحلة المصدر، دون تقديمها بوصفها وصفًا محدثًا للمعايير.

الكلمات المفتاحية: الويب الدلالي، الممارسات الجيدة، مجموعات البيانات، البيانات الوصفية، إعادة الاستخدام

مقدمة:

في مايو 2016، يكون قد مضى خمسة عشر عامًا على نشر (Berners-Lee) و (Hendler) و (Lassila) (2001) العمل الذي عرض أهم أهداف الويب الدلالي ومكوناته. وقد ذكر دون تفصيل بعض التقنيات التي أسهمت لاحقًا في تحديد بنيته العامة ومبادئه. وكان الانتقال من نشر الوثائق إلى التعبير عن معناها الخيط الناظم لتطوير معايير وتقنيات تستهدف تمثيل المعرفة على الويب. وشكل التشغيل البيني التركيبي باستخدام (XML) والتشغيل البيني الدلالي باستخدام (RDF) و (OWL) التقنيات الرئيسة التي بُني عليها. وخلال هذه السنوات بينت الممارسة أن تقنيات نشر البيانات الناضجة لا تكفي؛ فانتشار الويب الدلالي يعتمد أيضًا على إمكانات تطبيق البيانات واستثمارها، وعلى إرشادات واضحة دقيقة لعملية نشرها.

يهدف فريق (W3C) للممارسات الجيدة للبيانات على الويب (DWBP) إلى بناء منظومة للبيانات المفتوحة بتيسير التفاهم بين المطورين وناشري البيانات. ويتحقق ذلك بتحسين اتساق البيانات وتقديم الإرشاد المناسب للناشرين. ويركز الفريق على مسارين: تحديد ممارسات جيدة للتنفيذ التقني أثناء النشر؛ وتحديد مفردات لتمثيل مؤشرات جودة مجموعات البيانات واستخدامها.

فهم مجموعات البيانات:

تطور مفهوم مجموعة البيانات (dataset) مع الممارسة وظهور حاجات جديدة للنشر. ومن علامات نضج تقنيات الويب الدلالي أن الاهتمام لم يعد منصبًا بقدر كبير على حلول التحرير والنشر التقنية وحدها، بل على استخدام البيانات المنشورة وتراخيصها وموثوقيتها وآليات نشرها وإعادة استخدامها. ولذلك يجب أن يستند إعداد مجموعات البيانات وصيانتها إلى نهج مستقل عن أوجه استثمارها اللاحقة المحتملة. وتحاط المجموعة المنشورة على الويب بعناصر تقنية: بنية الويب نفسها، والمعايير والصيغ و (RDF) و (OWL)، والمفردات المستخدمة لتمثيل البيانات وآليات الوصول إليها. وإلى جانب قيم البيانات، يجب أن تضم المجموعة بيانات وصفية عن المحتوى والبنية والمنشأ ومؤشرات الجودة ووسائل الوصول والاستخدام. وتُنظم قيم البيانات في توزيعات مختلفة تستجيب لاحتياجات مختلفة للاستخدام أو الاستهلاك، ويتضمن كل توزيع معلومات عن صيغته وبنيته وترخيص استخدامه وغيرها.

تتجاوز هذه الجوانب النمذجة والتمثيل، مع بقائهما أساسيين. وبالأهمية نفسها ينبغي إدراج دورة حياة مجموعة البيانات في بيئة البيانات المفتوحة المترابطة (LOD) ضمن مفهومها. فلا تكفي النمذجة والنشر والتحديث الدوري؛ بل يلزم أيضًا (Auer et al., 2012): الربط بمجموعات أخرى؛ وإعداد مفردات مضبوطة للتصنيف؛ والإثراء؛ وتحليل الجودة؛ والتخطيط للاستثمار بآليات اكتشاف البيانات أو استخراجها؛ والاستعلام والبحث والتصفح.

التحديات والممارسات الجيدة:

يقوم نشر البيانات على الويب على التشغيل البيني باستخدام المعايير. ومن هذا المنظور تُعد الصيغ وآليات التعريف والوصول وإعادة استخدام المفردات وتطبيق البيانات الوصفية عناصر أساسية لتحقيق مستويات مناسبة من الصياغة الرسمية. ومع الوقت ظهرت تحديات تخص خصوصية البيانات المنشورة وأمنها والتحقق من جودتها ومنشئها. ويجب أن تشمل سياسات النشر تراخيص تحدد إمكانات الاستثمار. وأصبحت إتاحة إصدارات مختلفة من المجموعات وإثراؤها وحفظها وصيانتها خصائص ملازمة لجودتها.

يعد فريق (DWBP) وثيقة ممارسات جيدة تحدد هذه التحديات بدقة (W3C, 2016a)، انطلاقًا من متطلبات حالات الاستخدام (W3C, 2015a): استخدام مفردات البيانات لزيادة التشغيل البيني الدلالي، بإعادة استخدام المفردات واختيار مستويات مناسبة للنمذجة والصياغة الرسمية؛ وضمان خصوصية البيانات الحساسة وأمنها بتحديد الأجزاء غير المتاحة للعامة وإبلاغ المستخدمين بأسباب القيود.

واستخدام معايير متعددة لإتاحة البيانات بصيغ تزيد مرونة إعادة استخدامها؛ واجتذاب المستخدمين والمستهلكين ببيان كيفية الاتصال بالناشرين ونشر نتائج التغذية الراجعة بيانات مفتوحة؛ وإثراء البيانات بمجموعات وبيانات وصفية وقواعد بيانات وموارد إلكترونية أخرى؛ وتقديم بيانات وصفية يقرأها البشر والآلات، سواء أكانت وصفية أم بنيوية أم متخصصة، بمصطلحات من مفردات معيارية؛ وبيان ترخيص المجموعة ليعرف المستهلكون حدود إعادة الاستخدام وإمكاناته؛ وإضافة معلومات المنشأ والجودة لتقييم الموثوقية والثقة؛ وإدارة الإصدارات إدارة مناسبة مع تجنب تغييرات واجهات (API) أو خدمات الوصول التي تُلزم العملاء الحاليين بتعديل تطبيقاتهم.

واستخدام معرفات موارد (URI) دائمة لتعريف المجموعة كاملة ومواردها المفردة وإصداراتها؛ وتوفير آليات متعددة للوصول إلى البيانات المحدثة في الوقت الحقيقي قدر الإمكان، عبر تنزيل المجموعة كاملة بصيغة معيارية، وكذلك عبر خدمات الويب وواجهات (API) موثقة توثيقًا مناسبًا.

وانطلاقًا من هذه التحديات تُقترح 32 ممارسة جيدة، مجمعة حول فوائد: الفهم، بتحسين فهم البشر لبنية البيانات وطبيعتها ومعناها وللمجموعة وبياناتها الوصفية؛ والمعالجة، بتمكين الآلات من معالجتها والتصرف فيها آليًا؛ والاكتشاف، باكتشاف الآلات المجموعات أو بيانات محددة داخلها؛ وإعادة الاستخدام، بزيادة إمكاناته لدى فئات مختلفة من المستهلكين؛ والثقة، بتحسين ثقتهم بالمجموعة؛ وقابلية الربط، بإنشاء روابط بين الموارد على مستوى المجموعات والبيانات المفردة؛ والوصول، بإتاحة البيانات المحدثة للبشر والآلات بصيغ ومستويات تفصيل متعددة؛ والتشغيل البيني، بتيسير التوافق بين الناشرين والمستهلكين.

تُوصف كل ممارسة تفصيلًا بتبريرها وأهدافها ومقترحات تنفيذ عملية وأمثلة وآليات اختبار، مع أدلتها في وثيقة المتطلبات وحالات الاستخدام.

وهذا نهج أكثر تطورًا من المقترح السابق في مجال (LOD) (W3C, 2014). كما تُطبق هذه الممارسات مباشرة في مشروعات البيانات المفتوحة ضمن سياقات محددة، مثل الإدارات العامة (Pastor-Sánchez, 2014). ويعمل الفريق نفسه على مجموعتين من المفردات: أولاهما لتمثيل جوانب جودة مجموعات البيانات ومؤشراتها ومقاييسها (W3C, 2015b)، لتوسيع وظائف مفردات (DCAT) السابقة (W3C, 2014b)؛ والثانية في مرحلة مبكرة جدًا من مسودة العمل لتمثيل استخدامات المستهلكين لمجموعة البيانات (W3C, 2016b).

توسيع رؤية الويب الدلالي:

مر الويب الدلالي، كسائر التقنيات الناشئة، بدورة تطور ارتبطت خلالها التوقعات بإنجازات محددة. وخلال السنوات الخمس أو الست الماضية أشار مؤلفون ((Heath)؛ (Bizer)، 2011، الفصل 6؛ (Herman)، 2011؛ (Saorín)؛ (Peset)؛ (Ferrer-Sapena)، 2013) إلى أن الجوانب التقنية لا تكفي لضمان نجاح التقنية وبقائها.

نحن إذن أمام مقترحات تتجاوز التقنية إلى آليات نشر البيانات والوصول إليها وإعادة استخدامها. وقد كشفت الممارسة صعوبات محددة، فكان لا بد من إرشادات وممارسات جيدة توجه مهام النشر وتحسنها. ومع ذلك فالويب منظومة تقنية معقدة دائمة التطور؛ ومهما أفادت أعمال (DWBP) الآن، ستظهر مشكلات وتقنيات وتطبيقات جديدة، مما يستلزم مراجعة مستمرة للإرشادات. والويب الدلالي، الذي كان الإلمام بتقنياته سهلًا نسبيًا قبل سنوات، سيقتضي فهمه وتطبيقه الصحيحان تجاوز الجوانب التقنية والإلمام العميق بجوانب أخرى من دورة حياة البيانات على الويب.

المراجع:

- Auer, Sören; Bühmann, Lorenz; Dirschl, Christian; Erling, Orri; Hausenblas, Michael; Isele, Robert; Lehmann, Jens; Martin, Michael; Mendes, Pablo N.; Van-Nuffelen, Bert; Stadler, Claus; Tramp, Sebastian; Williams, Hugh (2012). "Managing the lifecycle of linked data with the LOD2 stack". En: 11th Intl semantic web conf, Boston, MA, USA, November 11-15, Proceedings, Part II. http://dx.doi.org/10.1007/978-3-642-35173-0_1

- Berners-Lee, Tim; Hendler, James; Lassila, Ora (2001). "The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities". Scientific American, v. 284, n. 5, pp. 28-37. http://goo.gl/aQH1S0
- Heath, Tom; Bizer, Christian (2011). Linked data: evolving the Web into a global data space. San Raphael, CA: Morgan & Claypool Publishers. ISBN: 978 1608454310 http://linkeddatabook.com/book
- Herman, Ivan (2011). Semantic web adoption and applications. http://www.w3.org/People/Ivan/CorePresentations/ Applications/Applications.pdf
- Pastor-Sánchez, Juan-Antonio (2014). "Aspectos prácticos para proyectos de datos abiertos en las administraciones públicas". Anuario ThinkEPI, v. 8, pp. 313-317.
- Saorín, Tomás; Peset, Fernanda; Ferrer-Sapena, Antonia (2013). "Factores para la adopción de linked data e implantación de la web semántica en bibliotecas, archivos y museos". Information research, v. 18, n. 1, paper 570. http://InformationR.net/ir/18-1/paper570.html
- W3C (2014a). Best practices for publishing linked data. W3C Working Group, Note 09 January. https://www.w3.org/TR/2014/NOTE-ld-bp-20140109
- W3C (2014b). Data catalog vocabulary (DCAT). W3C Recommendation, Note 16 January. http://www.w3.org/TR/vocab-dcat
- W3C (2015a). Data on the web best practices, use cases, & requirements. W3C Working Group, Note 24 February. http://www.w3.org/TR/2015/NOTE-dwbp-ucr-20150224
- W3C (2015b). Data on the web best practices: Data quality vocabulary. W3C Working Draft, 17 December. https://www.w3.org/TR/2015/WD-vocab-dqv-20151217
- W3C (2016a). Data on the web best practices. W3C Working Draft 12 January. http://www.w3.org/TR/2016/WD-dwbp-20160112
- W3C (2016b). Data on the web best practices: Dataset usage vocabulary. W3C Editor's Draft 26 January 2016. http://w3c.github.io/dwbp/vocab-du.html