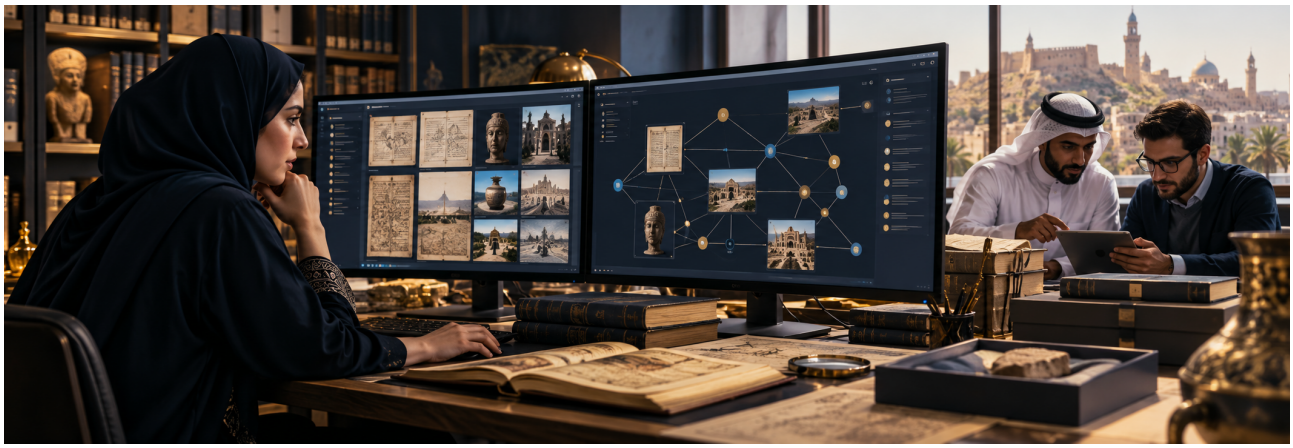
وصف مجموعات البيانات على الويب: مفردات تعزز الاكتشاف والثقة

خوان أنطونيو باستور سانشيث



الملخص:

يعرض المقال مفردات وأنطولوجيات تساعد وصف مجموعات البيانات المنشورة على الويب، وتزيد اكتشافها وإعادة استخدامها والثقة بها. ويشرح الفرق بين المجموعة وإصداراتها وتوزيعاتها ورسوم (RDF)، وأهمية المعرّفات المستقرة وحفظ تاريخ التغييرات. ثم يتناول (VoID) لوصف البيانات المترابطة و(DCAT) لبناء الفهارس، و(PROV-O) و(PAV) لتمثيل المنشأ وأدوار الفاعلين، و(DQV) لقياسات الجودة، و(DUV) للاستشهاد والاستخدام وملاحظات المستخدمين. ويبرز تكامل المحتوى والبنية والوصول والترخيص والجودة ضمن البيانات الوصفية، وإعادة استخدام المفردات القائمة لدعم التشغيل البيني. كما يناقش دور ملاحظات المستخدمين في التحسين، والحاجة إلى منصات وواجهات تيسر الاستفادة الفعلية. وتبقى توصيفات المعايير وآفاق تطورها مرتبطة بمرحلة المصدر، مع الحفاظ على ملاحظاته ومواضع عدم الاكتمال كما وردت.

الكلمات المفتاحية: مجموعات البيانات، البيانات الوصفية، الويب الدلالي، (DCAT)، التشغيل البيني

مقدمة:

ساعد تطور تقنيات الويب الدلالي ونضجها على تحديد تطبيقاتها الملموسة، وأتاح تنفيذ أدوات لنشر البيانات المهيكلة على الويب. وتستند إتاحة هذه البيانات إلى مبدأ التشغيل البيني الدلالي، وهو هدف معايير مثل (RDF) و(OWL) و(SPARQL) وغيرها.

تستخدم آلاف المجموعات هذه التقنيات للنشر والوصول. والتشغيل البيني لا يقوم على التقنيات المذكورة وحدها، بل على تعريف المفردات أيضًا. ومن المبادرات المهمة هنا (LOV)، التي تتيح للمبرمج البحث عن مفردات قائمة قبل تعريف مفردات جديدة تلبي حاجات محددة لوصف الموارد وتمثيلها. أي إن إعادة الاستخدام لا تقتصر على مجموعات البيانات، بل تشمل المفردات.

وتؤكد مجموعة عمل (Data on the Web) في (W3C) ذلك في وثيقة أفضل الممارسات (Farias-Lóscio; Burle; Calegari, 2016): إعادة استخدام المفردات تزيد التشغيل البيني للبيانات. ولذلك ينبغي أن تحتوي المجموعة على معلومات إضافية تتجاوز بياناتها نفسها؛ أي بيانات وصفية عن المجموعة، تشمل الجوانب الوصفية والبنيوية والارتباط بمجموعات أخرى، وكيفية الاستخدام، والمعايير والمؤشرات المتصلة بالجودة.

ما المقصود بمجموعة البيانات:

مجموعة البيانات أكثر من حاوية لعبارات (RDF). إنها وحدة واصفة لذاتها تضم بيانات مهيكلة متاحة بصيغ عدة للتنزيل الكامل أو الوصول الانتقائي. ويجب أن تراعي إدارة إصداراتها التحديثات عبر الزمن. ويلزم أيضًا مفهوم التوزيعة (distribution)، الذي ييسر استخدام المجموعة لفئات المستفيدين أو مستهلكي البيانات وفق حاجاتهم.

قد يفضل بعضهم تنزيل المجموعة كاملة بصيغة (RDF/XML)، ويفضل آخرون نقطة وصول (SPARQL Endpoint)، وقد يفضل غيرهم واجهة (API). ويمكن أن تكون للمجموعة إصدارات متتابعة، ولكل إصدار توزيعة أو أكثر.

ينبغي أن تقدم البيانات الوصفية معلومات عن المجموعة وإصداراتها وتوزيعاتها المختلفة، لتيسير الثقة وإعادة الاستخدام. وإلى جانب وصف المحتوى والبنية، يجب بيان كيفية الوصول ومؤشرات الجودة والمنشأ وترخيص الاستخدام.

كذلك يجب نشر جميع الموارد، سواء المجموعة كاملة أو عناصرها، بمعرّفات (IRI) مستقرة حتى يمكن الإحالة إليها. وعلى الناشرين مراعاة الإصدارات الجديدة، وإدراك أن حذف البيانات بلا بيان ممارسة غير جيدة؛ بل يلزم توضيح أن المورد حذف أو أُرشف، وتقديم المعلومات المناسبة. وهنا تبرز أهمية رسم (RDF) واستخدام معرّفات (IRI).

ليست موارد (RDF) العناصر الوحيدة التي يمكن تعريفها بواسطة (IRI)؛ فالمجموعة كاملة تعرّف بالطريقة نفسها. ومجموعة البيانات، في الحقيقة، مجموعة من رسوم (RDF).

يرتبط كل رسم في المجموعة، باستثناء واحد، بمعرّف (IRI)؛ وتسمى هذه «الرسوم المسماة» (named graphs). أما الرسم بلا (IRI) فيعرف بالرسم الافتراضي (default graph) (Bizer, 2007, p. 62).

وترتبط عبارات (RDF) برسوم مسماة محددة؛ لذلك يتسع مفهوم الثلاثيات <الموضوع><المحمول><الكائن> إلى الرباعيات <الموضوع><المحمول><الكائن><الرسم>. وتنتمي العبارة غير المرتبطة برسم مسمى إلى الرسم الافتراضي (Carothers, 2014).

قد يبدو الشرح تقنيًا، لكنه ضروري لفهم إدارة إصدارات المجموعات. ومع عدم وجود طريقة معيارية بعد، تُتّبع غالبًا ممارسة تتمثل في أن تمثل عبارات الرسم الافتراضي أحدث إصدار، بينما ترتبط عبارات الإصدارات السابقة برسوم مسماة. ويلزم حفظ الإصدارات السابقة ضمن خطة للصون الرقمي؛ إذ يتيح ذلك تدقيق المجموعة وتقييمها ويشكل مؤشر ثقة.

وصف مجموعات البيانات:

يلزم وصف جوانب من المحتوى والبنية، وهناك نهجان: إضافة بيانات وصفية ذاتية إلى البيانات نفسها؛ وإدارة فهرس للمجموعات لأغراض الضبط والبحث.

تستخدم (VoID)، أي مفردات مجموعات البيانات المترابطة (Vocabulary of Interlinked Datasets)، عادة للمهمة الأولى. وتهدف إلى وصف المجموعة وروابط (RDF) التي تصل مجموعات مختلفة (Alexander et al., 2011; Cyganiak et al., 2011). وهذه روابط يكون موضوعها وكائنها موصوفين في مجموعتين مختلفتين.

وتستخدم عادة مصطلحات (Dublin Core) أو (FOAF) لوصف معرّفات المجموعة وناشرها ومصدرها وتواريخ إنشائها ونشرها وتعديلها وغيرها.

وتتيح (VoID) أصنافًا وخصائص لتحديد المفردات المستخدمة، والتقسيمات، أي المجموعات الفرعية، والأصناف والخصائص والمعلومات الإحصائية عن البنية. كما تحدد آليات الوصول، مثل التنزيل ونقطة (SPARQL) والموارد التي توفر منافذ الوصول وأمثلة الموارد. وتصف أيضًا روابط المجموعات، بتحديد المجموعات المتصلة ونوع الروابط وعددها.

أما لبناء فهارس المجموعات فتُستخدم (DCAT)، أي (Data Catalog Vocabulary). وهي مفردات لوصف المجموعات في فهارس قابلة للتشغيل البيني، بحيث تستهلك التطبيقات أوصاف مجموعات من فهارس عدة معًا (Maali; Erickson, 2014).

ويمكن باستخدام (DCAT) تعريف فهارس سجلات تصف المجموعات. ومن جوانبها المهمة الإحالة إلى ناشري المجموعات داخل الفهرس نفسه، واستخدام مفردات (SKOS) لتصنيف المجموعات بحسب الموضوع.

المنشأ:

مجموعات البيانات كيانات معلوماتية ينتجها ناشر، وقد تعكس أنشطة محددة. ويقصد بالمنشأ هنا المعلومات عن المنظمات والأنشطة والأشخاص الذين ينتجون المجموعة. وتستخدم لتقييم الجودة والموثوقية والثقة. وتتيح أنطولوجيا (PROV-O) تمثيلها (Lebo; Sahoo, 2013).

وتعرّف (PROV-O) المجموعات كيانات تنتجها أنشطة تجري خلال مدة زمنية وترتبط بدورها بفاعل. وتنتظم أصنافها وخصائصها في ثلاثة مستويات من التفصيل.

أصناف وخصائص البداية: تشكل قاعدة (PROV)، وتنتج أوصاف منشأ بسيطة تقتصر على الكيانات والفاعلين والأنشطة.

الأصناف والخصائص الموسعة: توفر مصطلحات إضافية لوصف أدق للمنشأ، بما فيه الأشخاص والفاعلون البرمجيون والمجموعات وأوصاف المنشأ والمواقع.

الأصناف والخصائص المؤهلة: تصف بدقة كبيرة التأثير بين الكيانات والفاعلين والأنشطة، والمراجعات والتفويضات والارتباطات بين الفاعلين والكيانات المشتقة وغيرها.

قد تكون (PROV-O) معقدة جدًا للاستخدام أحيانًا. وتقدم أنطولوجيا (PAV) نهجًا أبسط ومكملًا في الوقت نفسه؛ إذ تخصص بعض مصطلحات (PROV-O) و(Dublin Core) لوصف التأليف وتنقيح المحتوى والإنشاء الرقمي للموارد الإلكترونية. ومع أنها «أنطولوجيا خفيفة»، توفر مصطلحات تميز أدوار الفاعلين تجاه المورد، كالإنشاء والاستيراد والاشتقاق والتنقيح والاسترجاع وإدارة الإصدارات (Ciccarese; Soiland-Reyes, 2015).

الجودة:

تقاس جودة المجموعة وفق خصائص محددة للمجموعة وتوزيعاتها. ومفردات جودة البيانات (DQV) امتداد لـ(DCAT) لتمثيل أمور مثل تكرار التحديث، وقبول التعديلات المقترحة من المستخدمين، والتزام الناشرين بالمحافظة على المجموعة وصونها (Albertoni; Isaac, 2016). ويمكن أن تساعد المبرمجين على الثقة بالبيانات التي تستخدمها تطبيقاتهم.

وتعرف (DQV) خمسة أنواع من المعلومات، بأصنافها وخصائصها، للتعبير عن جودة المجموعة أو التوزيعة: الإفادات والدرجات والتقديرات وشهادات الجودة؛ والمطابقة للمعايير المستخدمة؛ والسياسات أو الاتفاقات التي تحكمها معايير جودة البيانات؛ ومقاييس الجودة النوعية أو الكمية؛ والكيانات الداخلة في منشأ المجموعة أو التوزيعة.

وتعرّف (DQV) قياسات الجودة انطلاقًا من أبعاد ومقاييس؛ فالأبعاد تشير إلى خصائص محددة، والمقاييس تحدد إجراءات قياس الأبعاد.

ومن أهم جوانبها آلية اشتقاق توصيفات الجودة. فهي تنطلق من إمكان اشتقاق معلومات الجودة من معلومات أخرى؛ لذلك تتيح اشتقاق توصيفات لجودة مجموعة أو توزيعة من معايير أو قياسات موجودة أو معرفة سابقًا.

الاستخدام والاستشهاد وملاحظات المستخدمين:

يمكن للمستهلكين استخدام البيانات المنشورة على الويب بطرائق متعددة. ومن المفيد للناشرين معرفة تجربة استخدامها. ويحتاج المستخدمون بدورهم إلى آليات للاستشهاد بالمجموعات أو معرفة أدوات الوصول المتاحة.

ولهذا عُرّفت مفردات استخدام البيانات (DUV)؛ لتوفر طريقة للاستشهاد بالبيانات، وبيانات وصفية عن كيفية الوصول إلى المجموعات وتوزيعاتها واستخدامها، وتمثيل تعليقات مستهلكي البيانات وتجاربهم (Farias-Lóscio; Stephan; Purohit, 2016).

بنية (DUV) قائمة على وحدات، وتقدم ثلاثة نماذج فرعية: الاستشهاد والاستخدام وملاحظات المستخدمين.

نموذج الاستشهاد: المجموعات والتوزيعات أشكال من الوسائط الإلكترونية يمكن للمستهلكين الاستشهاد بها باستخدام بيانات مرجعية ببليوجرافية أساسية يقدمها الناشرون. ويمكن للناشرين أيضًا إرفاق مراجع ببليوجرافية توفر معلومات إضافية للمستهلكين، بل ومراجع يقدمها المستهلكون أنفسهم.

نموذج الاستخدام: من المفيد تتبع كيفية استخدام المستهلكين للبيانات. ويمكن أن يحسن ذلك فائدتها، بوصف طرائق استخدامها وإدراج الوصف ضمن بيانات وصفية يقدمها الناشر إلى مجتمع المستهلكين.

نموذج ملاحظات المستخدمين: يوفر وسيلة لمشاركة ملاحظات التقييم. وهي مهمة لحل مشكلات جودة البيانات المنشورة؛ فقد تختلف تجارب المستخدمين مع المجموعة نفسها. ويتيح هذا النموذج تمثيل سياق الاستخدام وملف المستخدم. وتتيح هذه الملاحظات أيضًا إيصال التصحيحات أو الاقتراحات إلى الناشر.

خاتمة:

أصبح نشر البيانات وفق معايير التشغيل البيني أمرًا مسلمًا به. ويمكن القول إن مرحلة تطوير التقنيات والنشر التجريبي للمجموعات قد جرى تجاوزها. وصار تطبيق البيانات المفتوحة المترابطة معلومًا، وتحديد الصلات والمطابقات وروابط (RDF) بين المجموعات ممارسة شائعة. وأصبحت أطر تطوير برمجيات (RDF) وتخزينها مجالًا مطروقًا.

ومع ذلك، ما زال مجال بحث معماريات التطبيقات لنشر البيانات والاستعلام عنها خصبًا. فتطبيق أفضل الممارسات، وإنشاء مفردات للبيانات الوصفية التي تشرح المحتوى والبنية والمنشأ والجودة والاستخدام، يتيح تحديد خصائص منصات النشر بدقة أكبر.

وفي المستقبل القريب، سنشهد نشاطًا مكثفًا في إنشاء مفردات مثل (DCAT) و(PROV-O) و(DQV) و(DUV) وإعادة صياغتها وتبسيطها ونشر استخدامها الفعلي.

بل يمكن تصور أفق أوسع؛ إذ يحتمل أن يصبح نشر البيانات ووصفها على الويب وفق أفضل الممارسات أمرًا مفترضًا. وعندئذ يكون المفتاح كيفية استخدامها عبر واجهات (API) وتطبيقات الأطراف الأخرى.

المراجع:

- https://www.w3.org/standards/semanticweb/ontology
- https://lov.okfn.org
- https://www.w3.org/TR/rdf11-concepts/#dfn-iri
- http://dublincore.org/documents/2010/10/11/dcmi-terms
- http://xmlns.com/foaf/spec/
- Albertoni, Riccardo; Isaac, Antoine (2016). Data on the Web best practices: Data quality vocabulary. W3C Working Group Note, 30 August 2016. https://www.w3.org/TR/2016/NOTE-vocab-dqv-20160830
- Alexander, Keith; Cyganiak, Richard; Hausenblas, Michael; Zhao, Jun (2011). Describing linked datasets with the VoID vocabulary. W3C Interest Group Note, 03 March 2011. http://www.w3.org/TR/2011/NOTE-void-20110303
- Bizer, Christian (2007). Quality-driven information filtering in the context of Web-based information systems. Tesis Doctoral, Freie Universitat Berlin. http://www.diss.fu-berlin.de/diss/servlets/MCRFileNodeServlet/FUDISS_derivate_000000002736/
- Carothers, Gavin (2014). RDF 1.1 N-Quads: A line-based syntax for RDF datasets. W3C Recommendation, 25 February 2014. https://www.w3.org/TR/2014/REC-n-quads-20140225
- Ciccarese, Paolo; Soiland-Reyes, Stian (2015). PAV - Provenance, authoring and versioning. http://pav-ontology.github.io/pav ; http://purl.org/pav/2.3
- Cyganiak, Richard; Zhao, Juan; Alexander, Keith; Hausenblas, Michael (2011). "Vocabulary of Interlinked Datasets (VoID)". Deri Vocabularies. http://vocab.deri.ie/void

- Farias-Lóscio, Bernadette; Burle, Caroline; Calegari, Newton (2016). Data on the Web best practices. W3C Candidate Recommendation, 30 August 2016. https://www.w3.org/TR/2016/CR-dwbp-20160830
- Farias-Lóscio, Bernadette; Stephan, Eric G.; Purohit, Sumit (2016). Data on the Web best practices: Dataset usage vocabulary. W3C Working Group Note, 30 August 2016. https://www.w3.org/TR/2016/NOTE-vocab-duv-20160830
- Lebo, Timothy; Sahoo, Satya (2013). PROV-O: The PROV ontology. W3C Recommendation, 30 April 2013. http://www.w3.org/TR/2013/REC-prov-o-20130430/
- Maali, Fadi; Erickson, John (2014). Data catalog vocabulary (DCAT). W3C Recommendation, 16 January 2014. http://www.w3.org/TR/2014/REC-vocab-dcat-20140116