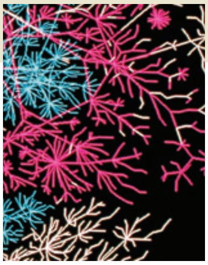


_computational BIOLOGY

PERSPECTIVE

From XML to RDF: how semantic web technologies will change the design of 'omic' standards

Xiaoshu Wang, Robert Gorlitsky & Jonas S Almeida

With the ongoing rapid increase in both volume and diversity of 'omic' data (genomics, transcriptomics, proteomics, and others), the development and adoption of data standards is of paramount importance to realize the promise of systems biology. A recent trend in data standard development has been to use extensible markup language (XML) as the preferred mechanism to define data representations. But as illustrated here with a few examples from proteomics data, the syntactic and document-centric XML cannot achieve the level of interoperability required by the highly dynamic and integrated bioinformatics applications. In the present article, we discuss why semantic web technologies, as recommended by the World Wide Web consortium (W3C), expand current data standard technology for biological data representation and management.

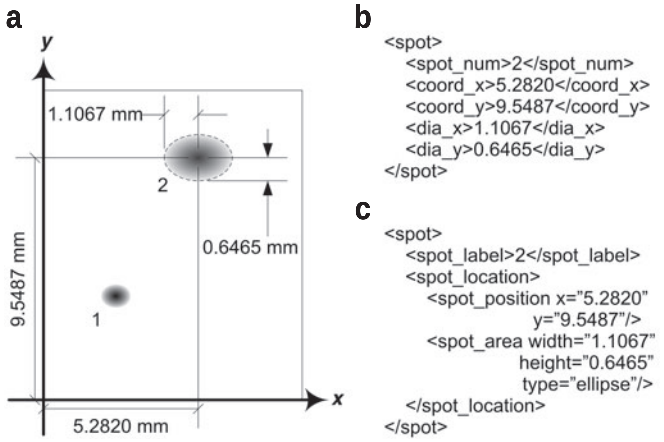
Developing a data standard addresses two major concerns. The first is the content—what should be standardized, and the second is the methodology—how the standardization should be formatted. Most discussions about data standardization in life sciences have been directed almost exclusively to the former[1,2]. But the choice of the standard technology in fact conditions not only how the data are accessed but more importantly determines whether crossdiscipline content can be merged to allow systemic integration, which is the critical issue for omic-level studies of biological organisms[3,4].

Furthermore, a data standard is more than just a medium to uniform data representation. By laying out the overall structure of relationships of the encoded data, a data standard will effectively define a schema for a particular area of domain knowledge. In this account, a data standard resembles the basic 'form of intuition', which, in a Kantian interpretation, conditions human perception during knowledge generation[5]. In addition, a data standard, once accepted, will become the *lingua franca* for the respective community. Indeed, linguists have long postulated that language is not a mere label but the very origin of thought[6]. As a recent study on the numerical cognition of a Brazilian tribe has forcefully demonstrated, human cognition itself is constrained by the language formalism[7]. Finally, standardization in the information age has a unique characteristic in that it is often carried out prior to or in parallel with technology development[8–10]. As history has demonstrated that a standard developed with incomplete knowledge could indeed hamper innovation[11], additional care must be taken to ensure that a designed data standard can evolve and adapt to a changing paradigm.

The purpose of this article is therefore to discuss how the above issues affect the choice of methodologies to establish data standards. More specifically, the article aims at discussing the need and options to go beyond the currently preferred choice of XML as a standard technology[12] to represent biological data.

The limitations of XML

To help understand the problem in detail, a hypothetical two-dimensional gel electrophoresis (2DE) gel experiment was devised (Fig. 1a) along with XML fragments for describing the location and shape of spot 2 in two markup languages—annotated gel markup language (AGML)[13] and human proteome markup language (HUP-ML)[14] (Fig. 1b,c). The difference between the two XML formats shows that compatibility cannot be achieved by XML alone because the language can be used in more than one way to encode the same information. It may be argued that the compatibility issue would have never occurred if the two parties had agreed on a single standard. This is undoubtedly true. But the question is: 'how can it be achieved in XML?' In any scientific discipline data relationships are bound to change with the development of new



```
<spot>
  <spot_num>2</spot_num>
  <coord_x>5.2820</coord_x>
  <coord_y>9.5487</coord_y>
  <dia_x>1.1067</dia_x>
  <dia_y>0.6465</dia_y>
</spot>
```

```
<spot>
  <spot_label>2</spot_label>
  <spot_location>
    <spot_position x="5.2820"
                   y="9.5487"/>
    <spot_area width="1.1067"
               height="0.6465"
               type="ellipse"/>
  </spot_location>
</spot>
```

Figure 1 A hypothetical 2DE example. (**a**) An artificially created 2DE gel with two 'spots'. (**b**,**c**) The description of the location and shape of the second spot is shown in AGML (**b**) and HUP-ML (**c**). Note that the two XML formats differ significantly in syntax and neither schema explicates the assumed coordinate system and the axis-aligned elliptic shape of spot in **a** that are necessary to make the XML codes meaningful.

Department of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina, 135 Cannon St. Suite 303, Charleston, South Carolina 29403-5720, USA. Correspondence should be addressed to J.S.A. (almeidaj@musc.edu).

Published online 7 September 2005; doi:10.1038/nbt1139

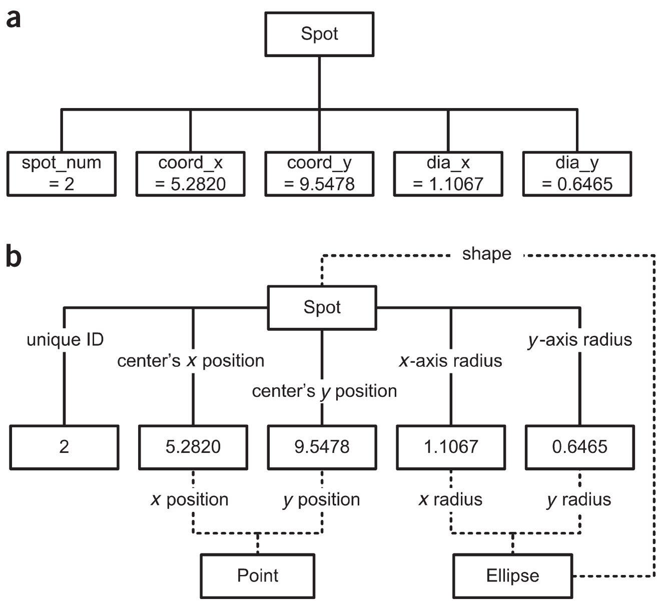


Figure 2 Data relationships for a spot on a 2DE gel and its XML representation. (**a**) The data model of AGML schema for a 2DE spot. Only the relevant elements discussed in the text are shown. (**b**) The data semantics of a 2DE spot. The semantics that can be mapped to the AGML structures are shown in solid lines, whereas those that can not be mapped but are implicitly assumed in AGML are shown in dotted lines. All these semantics can be made explicit by the RDF representation as discussed in **Box 1** and modeled in **Figure 4**.

experimental methods. When such a change occurs, the standard must adjust to reflect the newly established relationship. Unfortunately, it is very difficult to define such an adaptive standard in XML. For instance, the concept of a virtual gel that underlies AGML—a catalog gel generated from a set of aligned real gels—is not supported by HUP-ML. To extend HUP-ML to support the concept would require an additional data item to be included in the original scheme. The task would appear at first to be a straightforward exercise because simply adding an optional attribute, such as “virtualGel_id”, to the original <spot> construct would suffice.

But such a simple request demands a nontrivial solution in reality. First, the scheme of HUP-ML does not allow any extension of the vocabulary beyond the original specifications. This rigid requirement is not a design failure of HUP-ML, but instead it reflects the nature of XML. By restricting what type of data can and cannot be in what places, an XML-encoded message can be validated to ensure correct software operation. Of course, techniques such as wildcard or substitution group can be used to equip a schema with flexible extensibility. Use of these techniques, however, unrealistically requires the schema designer to anticipate all future developments of the experimental method. Even so, because no rule can be specified to restrict the manner of extension, separately developed applications are very likely to develop different ‘dialect’ extensions[15]. What is worse, because XML-based applications depend on the correct document structure to operate properly, any structural change may potentially break the applications that support the original format. Hence, a simple extension will effectively create two different standards, defeating the original purpose of using the built-in flexibility to extend a common standard. An alternative solution is to group newly extended features into a new namespace. Such an approach avoids breaking the existing schema, but the newly extended feature is unlikely to be structurally cohesive with the existing ones. The <virtualGel_id> element, for instance, must be arbitrarily placed at a location that is not obviously related to <spot>. In this case, software design, instead of scheme design, becomes an integration project, and the incompatibility remains.

The difficulty of extending XML-based standards has prompted many standard designers to bulge their schemes to anticipate future developments. For instance, the gel-centric standards—both AGML and HUP-ML—have designed elements to accommodate the possible inclusion of mass spectrometry (MS) data. Conversely, mzXML[15], a standard developed mainly for encoding MS data, has designed a variable content holder for potential 2DE data. But because these standards differ, despite overlapping with each other in design philosophy, convention, techniques and even the required contents, merging 2DE data with MS data is even harder to achieve than merging the 2DE standards.

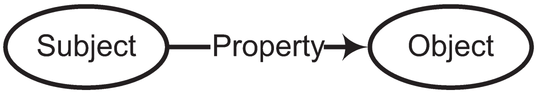


Figure 3 Graph model for an RDF statement. An RDF statement can be modeled as a DLG with resources (subject and object) as nodes and properties as the edge connecting from ‘subject’ to ‘object’.

To say the least, even if the above integration difficulties can be overcome so that all data standards can be unified into a single markup language, the resulting schema would be of no practical use. All data are inherently related with each other. To accommodate all possible relationships, the grand scheme will eventually reach a magnitude that is simply too complex to implement.

Where do the problems with XML originate?

The above problem originates from the limited expressiveness of the XML language. This claim may appear to contradict the often proclaimed ‘self-descriptive’ nature of XML. But XML, designed as a language for message encoding, is only self-descriptive about the following structural relationships: containment, adjacency, co-occurrence, attribute and opaque reference. All these relationships “are indeed useful for serialization, but are not optimal for modeling objects of a problem domain”[16]. For instance, the relationship between the <spot> and <coord_*> of AGML tags is no different from that between <spot> and <dia_*>. But a computer algorithm must nevertheless treat them differently to develop meaningful applications. To calculate the distance between two <spot>s, an algorithm shall use the value of <coord_*>, but to calculate the area of each <spot>, it shall retrieve the value of <dia_*> instead. This simple example illustrates that meaningful data exchange involves two levels of communication. The first is at the message level. At this level, data must be encoded and decoded in a standard format so that applications can know how to convert electronic bits into the data objects that a programming language can work with. The second level of communication is at the algorithmic level. At this level, the relationships between data objects must be explicitly specified so that applications can process the data accordingly.

XML is a language designed to standardize the communication at the message level. As shown in Figure 2a, the AGML schema describes a precise structural relationship between <spot> and its attributes. What appears to be missing is the description of the semantic relationships between nested content holders (Fig. 2b) that are required to invoke appropriate algorithms. Using XML alone at both levels requires a mapping of the domain knowledge to document structure. Considering that only a few types of relationships are specified in XML, the task is difficult, if not impossible, to achieve.

Box 1 Description of a 2DE spot in RDF

Document 1:

```
<rdf:RDF
  xmlns:rdf=”http://www.w3.org/1999/02/22-rdf-syntax-ns#”
  xmlns:cce=”http://www.charlestoncore.org/ontology/example#”>
  <cce:Spot rdf:about=”http://www.charlestoncore.org/ont/example/spot2”>
   <cce:shape>
     <cce:Ellipse>
       <cce:x-radius>1.1067</cce:x-radius>
       <cce:y-radius>0.6465</cce:y-radius>
       <cce:center>
         <cce:Point>
           <cce:x-position>5.2820</cce:x-position>
           <cce:y-position>9.5478</cce:y-position>
         </cce:Point>
       </cce:center>
     </cce:Ellipse>
    </cce:shape>
  </cce:Spot>
</rdf:RDF>
```

Document 2:

```
<rdf:RDF
  xmlns:rdf=”http://www.w3.org/1999/02/22-rdf-syntax-ns#”
  xmlns:cce=”http://www.charlestoncore.org/ontology/example#”
  xmlns:sup=”http://www.charlestoncore.org/ontology/supplement#”>
  <cce:Spot rdf:about=”http://www.charlestoncore.org/example/spot2”>
    <sup:virtualGel
  rdf:resource=”http://www.charlestoncore.org/ont/example/gel3”/>
  </cce:Spot>
</rdf:RDF>
```

The two independent RDF documents describe the same spot #2 (**Fig. 1a**). Document 1 describes its location and shape—the same information that the XML fragments of **Figure 1b**,**c** intended to encode. The content of document 1 corresponds to the graphic model (**Fig. 4**) composed in solid lines. Document 2 describes the virtual gel information (catalog information for spots that can be consistently found in a stack of aligned gels) for the same spot #2. Its content corresponds to the model graph (**Fig. 4**) composed in dotted lines. Some guidelines to understand the RDF/XML syntax are provided here. The URL links in this document are active and the reader is encouraged to try them. See also note at the end of the box. These two independent RDF documents describe a particular resource (http://www.charlestoncore.org/ont/example/spot2), whose nature is defined to be an (http://www.charlestoncore.org/ontology/example#Spot). The “#” signals a fragment identifier by the definition of URI. The simplest way to view this URI is to think it as a “Spot” defined in “http://www.charlestoncore.org/ontology/example”, which will lead to an RDF document. Within the retrieved document, another resource “http://www.charlestoncore.org/ont/example/” is specified by <rdfs:isDefinedBy>. It is important to understand the difference among the three types of documents encountered so far in this report. (i) The RDF documents presented in this Box. (ii) The document retrieved from “…/ontology/example” and (iii) the document retrieved from “…/ont/example”. Document (i) contains representations of particular data instances; document (ii) is the standard in RDF and document (iii) is a convenient description of (ii) using a natural language (English). The key difference between (ii) and (iii) is that the former is intended for machine, whereas the later is for human comprehension. Nevertheless, according to the specification of RDF, document (iii) is neither required to be retrievable, nor to be in a human comprehensible form. The lack of such a document, however, will make an open standard effectively closed. It is to the authors' belief that it is the best practice to make both (ii) and (iii) a requirement of practical implementations of RDF.

NOTE: All URIs formatted as URLs contain live documents to provide the reader with a interactive and syntactically consistent illustration. The contents of those documents are also provided as **Supplementary Notes**.

In its essence, a data interoperation problem is a communication problem and successful communication must use a language that is semantically transparent relative to what is communicated. As no single language has yet, or perhaps will ever, exist to establish the ‘universal truth’[17], any language is only capable of conveying a particular portion of human knowledge as machine-processible information[18]. The difficulties of using XML to exchange domain knowledge are therefore not so much because the language itself is flawed, which it is not[12], but because it is semantically underdetermined for the topic to be communicated.

Semantic web technologies

What is needed for solving the above interoperability issue is a knowledge-representation technology that can explicitly describe the data semantics. Such technology—the jointly named semantic web technologies has been recently endorsed by the W3C as the technology to promote data automation and reuse in the web (http://www.w3.org/2001/sw/).

The foundation semantic web technology is the resource-description framework (RDF). RDF, as its name suggests, is a system to describe resources. RDF has a very simple yet elegant data model that can be summed up in one sentence: everything is a resource that connects with other resources via properties. A resource, according to the RDF primer[19], "is anything that is identifiable by a uniform resource identifier (URI) reference". A property is also a resource but used to describe the relationship between resources.

The basic information unit in RDF is an RDF statement in the form of '(subject, property, object)'. Each RDF statement can be modeled as a graph comprising two nodes connected by a directed arc (Fig. 3). A set of such graphs can jointly form a directed labeled graph (DLG) that can in theory model most, if not all, domain knowledge. For instance, the RDF graph shown in Figure 4 can be used to describe the example "spot #2".

As a graph, the RDF model is oblivious to both syntax and semantics. But an RDF model can be serialized in the syntax of either XML[20], or N3 (ref. 21) or even a specialized graphical notation language such as DLG[2] (ref. 22). The semantics of an RDF model, on the other hand, are obtained via reference to RDF schema language (RDFS)[23] and ontology web language (OWL)[24]. RDFS and OWL are two other semantic web technologies. Both languages are layered on top of RDF to offer support for inference and axiom—two features that make semantic web technologies a departure from data representation toward knowledge representation[25].

Data and data standards in semantic web

To obtain a thorough comprehension of the semantic web technology, the same information provided in the earlier XML example (Fig. 1) has been described in RDF (document 1 in Box 1). Comparing RDF with XML reveals three important differences.

First is the use of data standards. A data standard in semantic web will be referred to as an ontology, which is a knowledge representation term defined as "a specification of a conceptualization"[26] or more specifically as "an engineering artifact to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary words"[27]. An ontology in this context is a dictionary, formulated in certain syntax, to embody concepts of a domain-specific knowledge. The RDF in 'document 1' uses one such user-defined ontology (http://www.charlestoncore.org/ontology/example). Unlike the namespaces in XML, which ultimately are unique character strings for grouping related concepts, the ontology URI in RDF must be retrievable. Following the ontology defined above, for example, will lead to an RDF document, in which the concepts and usage of "Spot", "Ellipse", "Point", "shape" and "center", among others, are defined.

The second difference of RDF is that the description of semantic relationship is explicit. Instead of using a combination of document structure and tag names to infer the shape of spot #2 as in XML, RDF explicitly states that the resource "http://www.charlestoncore.org/ont/example/spot2" is a "cce:Spot", whose "cce:shape" is an "cce:Ellipse", "cce:x-radius" is 1.1067 and "cce:y-radius" is 0.6465.

The third difference in RDF is that the unique identifier attribute used in XML is no longer needed. This is due to the fact that resource in RDF has a URI by definition. It is important to note that using a URI in RDF is fundamentally different from using unique identifiers in XML because the uniqueness of the former is ensured globally, whereas that of the latter is only guaranteed within a document. The document-centric view of XML makes it difficult to refer an embedded entity outside its XML document. For instance, how can "spot #2" be referred to outside of this article? Therefore, to make information cohesive in XML, all data have to be included within a single document. The situation in RDF is different because using URI makes the physical location of the statement irrelevant. For example, to supplement the virtual gel information about "http://www.charlestoncore.org/ont/example/spot2", another RDF (document 2 in Box 1) is sufficient.

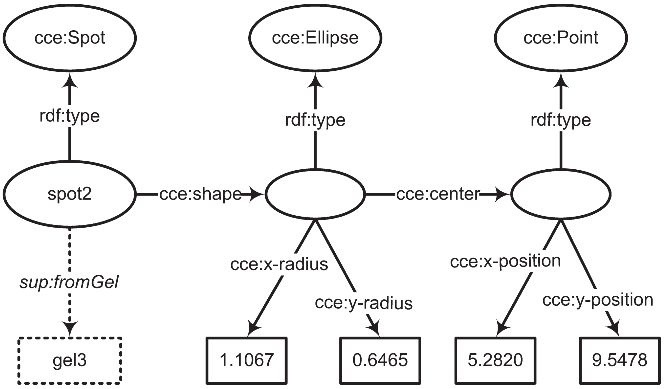


Figure 4 An RDF model for a spot on a 2DE gel. The graph in solid line illustrates an RDF model for a protein spot on a gel, and the graph in dotted line shows how to extend original model with an additional statement. The namespace "cce" refers to the example ontology defined in "http://www.charlestoncore.org/ontology/example", whereas "exp" refers to the supplement ontology defined in "http://www.charlestoncore.org/ontology/supplement". The "...spot2" and "...gel3" are shorthand for the URI of "http://www.charlestoncore.org/ont/example/spot2" and "http://www.charlestoncore.org/ont/example/gel3", respectively. The nodes without labels indicate blank nodes. To keep the graph as simple as possible, not all relationships are shown such as the domain and range of all properties of the example ontology. See **Box 1** for an illustration of two independent RDF documents using this model to provide information about the spot.

Why can RDF be helpful to omic approaches to biology?

Three distinct features of RDF make it very helpful to omic sciences. First, the data structure that endorsed the RDF is a DLG. Because adding nodes and edges to a DLG does not change the structure of any existing subgraph, RDF does not suffer the unpredictable extension-induced change in data structure that hampers the adaptability of the XML-based standard. Adding new information with new vocabularies to an existing resource is as easy as drawing a new node and connecting it to an existing graph (Fig. 4). Second, RDF has an open-world assumption in that it "allows anyone to make statements about any resource"[28]. Furthermore, RDF is monotonic in that new statements neither change nor negate the validity of previous assertions, making it particularly suitable in an academic environment, in which consensus and disagreement about the same resources have a useful coexistence that needs to be formally recorded. At last, all RDF terms share a global naming scheme in URI, making distributed data and ontologies possible.

The combined effect of global naming, universal data structure and open-world assumption is that resources exist independently but can be readily linked with little, if any, precoordination. For instance, the RDF in "document 2" (Box 1) not only provides additional information about spot #2, but it also uses a vocabulary "(http://www.charlestoncore.org/ontology/supplement#virtualGel)" that was not previously defined in "http://www.charlestoncore.org/ontology/example". The decoupled nature of RDF makes it a natural choice for defining an omic standard. The essence of omic science resides in its 'holistic' description of the subject of interest, and RDF makes it possible to connect all omic-specific data as a whole without necessarily turning them into a "whole".

Discussion

Just as any evolving new technology, RDF is not without issues. One particular problem of RDF is the vagueness of "resource" definition. When using a universal resource locator (URL)—instead of a URI—to represent resources of multiple dimensionalities, an "identity crisis" occurs[29]. The philosophical argument of what a URI represents is beyond the scope of this discussion[30]. In practice, the problem can be conveniently avoided by using the proposed life science identifier (LSID; http://www.omg.org/cgi-bin/doc?dtc/04-05-01). Because LSID is designed to couple a naming scheme with a data-retrieving framework, the design decision can be deferred to the implementation stage, when the owner of the resource can decide in what dimensionality the resource will be provided, or if any at all.

Of course, bristling alternative ontologies may emerge at the initial stage of ontological development for a particular scientific discipline. But, as a field matures, it is expected that the ontology usage will converge to the most efficient and comprehensive subset. The fact that RDF uses URI is in particular helpful in this regard. By assigning each concept a URI that can be globally referenced, RDF is immune to 'dialects'[15] that vexed XML-based standards. In RDF, whether an ontology becomes a 'standard' is mostly decided by its usefulness for a community. Opting for technologies that allow elected standards not only fits the natural progression of science, but that of human language as well[31].

It should be emphasized that, originated from knowledge representation, semantic web technologies are aimed at ultimately furnishing the current web with an inferencing engine. The usefulness of ontology is nonetheless independent of the availability of such an engine. First and foremost, the use of an ontology is to provide a lexicon. In this regard, RDF, by operating at semantic level, offers a uniform data representation medium that permits system interoperability through shared ontology[32].

Although the road to this vision is yet to be cleared, the life sciences community has already started moving in this direction. For instance, the Microarray Gene Expression Data Society (MGED) has started an Ontology working group (http://mged.sourceforge.net/ontologies/index.php) in an attempt to expand the concepts of MIAMI[33] from MGED-OM and MGED-ML[34] into RDF[35]. Projects have also been undertaken to express terms of Gene Ontology (GO) (http://www.geneontology.org/GO.format.shtml) and UniProt (http://www.isb-sib.ch/~ejain/rdf/) in RDF format. Last year, W3C sponsored the first workshop on life sciences where many topics and issues have been discussed[36]. Supporting tools for RDF, even if still limited and unstable when compared to its XML counterparts, are increasingly available (see http://www.w3.org/RDF/). What is now missing is a broader awareness of the fundamental XML conundrum and a clearer comprehension of the RDF technology among life scientists, such that they can participate more effectively in advancing the representation of their own domain expertise—a void this article hopes to assist filling.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS
This work was supported by the US National Heart, Lung, and Blood Institute (NHLBI) Proteomics Initiative through contract N01-HV-28181 to the Medical University of South Carolina, Principal Investigator. D. Knapp, and its bioinformatics core (core C, Principal Investigator J.S. Almeida) and mathematical modeling project (project 7, Principal Investigator E.O. Voit), as well as by its administrative center, separately funded by the same initiative to the same institution, Principal Investigator M.P. Schachte. The authors also acknowledge support by the training grant 1-T15-LM07438-01 "training of toolmakers for Biomedical Informatics" by the US National Library of Medicine of the National Institutes of Health, (NIH/NLM).

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/

1. Quackenbush, J. Data standards for 'omic' science. *Nat. Biotechnol.* **22**, 613–614 (2004).
2. Brazma, A. On the importance of standardisation in life sciences. *Bioinformatics* **17**, 113–114 (2001).
3. Zerhouni, E. Medicine. The NIH Roadmap. *Science* **302**, 63–72 (2003).
4. Check, E. NIH 'roadmap' charts course to tackle big research issues. *Nature* **425**, 438 (2003).
5. Kant, I. *Critique of Pure Reason*, 2nd revised edn. (Palgrave Macmillan, New York, 2003).
6. Whorf, B.L. Language, mind and reality. *Theosophist* **63**, 281–291 (1942).
7. Gordon, P. Numerical cognition without words: evidence from Amazonia. *Science* **306**, 496–499 (2004).
8. Cargill, C. Information Technology Standardization: Theory, Process And Organizations (Digital Press, Bedford, Massachusetts, 1989).
9. Krechmer, K. The fundamental nature of standard: technical perspective. *IEEE Commun. Mag.* **38**, 70 (2000).
10. Sherif, M. A framework for standardization in telecommunications and information technology. *IEEE Commun. Mag.* **39**, 94–100 (2001).
11. Farrell, J. & Saloner, G. Standardization, compatibility and innovation. *Rand J. Econ.* **16**, 70–83 (1985).
12. Barillot, E. & Achard, F. XML: a lingua franca for science? *Trends Biotechnol.* **18**, 331–333 (2000).
13. Stanislaus, R., Jiang, L.H., Swartz, M., Arthur, J. & Almeida, J.S. An XML standard for the dissemination of annotated 2D gel electrophoresis data complemented with mass spectrometry results. *BMC Bioinformatics* **5**, 9 (2004).
14. Kamijo, A. *et al.* HUP-ML: Human proteome markup language for proteomics database. *JMSSJ On-line* **51**, 542–549 (2003). http://db.wdc-jp.com/mssj/search/abst/200305/ms510542.html
15. Pedrioli, P.G. *et al.* A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **22**, 1459–1466 (2004).
16. Cover, R. XML and semantic transparency. *The Cover Pages*, published online 23 October 1998, revised 24 November 1998. http://www.oasis-open.org/cover/xmlAndSemantics.html
17. Spender, J. Pluralist epistemology and the knowledge-based theory of the firm. *Organ.* **5**, 233–256 (1998).
18. Galliers, R.D. & Newwell, S. Back to the future: from knowledge management to data management. in *Proceedings of the 9th European Conference on Information Systems 2001*, Bled, Slovenia, June 27–29, 2001, 609–615 (Moderna Organizacija, Kranj, Slovenia, 2001).
19. Manola, F. & Miller, E. RDF Primer. W3C Recommendation published online 10 February 2004. http://www.w3.org/TR/rdf-primer/
20. Beckett, D. RDF/XML Syntax Specification (Revised). W3C recommendation published online 10 February 2004. http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/
21. Berners-Lee, T. Primer: Getting into RDF & Semantic Web using N3. Published online 29 June 2005. http://www.w3.org/2000/10/swap/Primer.html
22. Wang, X. & Almeida, J.S. DLG^2 - A Graphical Presentation Language for RDF and OWL (v 2.0). Published online 10 August 2005. http://www.charlestoncore.org/dlg2/
23. Brickley, D. RDF Vocabulary Description Language 1.0: RDF Schema. W3C recommendation published online 10 February 2004**.** http://www.w3.org/TR/rdf-schema/
24. McGuinness, D.L. & van Harmelen, F. OWL Web Ontology LanguageOverview. W3C recommendation published online 10 February 2004. http://www.w3.org/TR/owl-features/
25. Davis, R., Shrobe, H. & Szolovits, P. What is a knowledge representation? *AI Magazine* **14**, 17–33 (1993).
26. Gruber, T. A translation approach to portable ontologies. *Knowledge Acquisition* **5**, 199–220 (1993).
27. Guarino, N. Formal Ontology and Information Systems, in: Formal Ontology in Information Systems (IOS Press, Amsterdam, Netherlands, 1998).
28. Klyne, G. & Carroll, J.J. (eds.) Resource Description Framework (RDF):Concepts and Abstract Syntax. W3C recommendation published 10 February 2004. http://www.w3.org/TR/rdf-concepts/
29. Clark, K.G. Identity crisis. *XML.com.* Published online 11 September 2002. http://www.xml.com/pub/a/2002/09/11/deviant.html
30. Berners-Lee, T. What do HTTP URIs identify? Published online 27 July 2002. http://www.w3.org/DesignIssues/HTTP-URI
31. Sole, R. Language: syntax for free? *Nature* **434**, 289 (2005).
32. Berners-Lee, T. & Hendler, J. Publishing on the semantic web. *Nature* **410**, 1023–1024 (2001).
33. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).
34. Spellman, P.T. *et al.* Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* 3, research0046 (2002).
35. Stoeckert, C.J. Jr., Quackenbush, J., Brazma, A. & Ball, C.A. Minimum information about a functional genomics experiment: the state of microarray standards and their extension to other technologies. *Drug Discov. Today*: *TARGETS* **3**, 159–164 (2004).
36. Summary Report**:** W3C Workshop on Semantic Web for Life Sciences. Published online 22 November 2004. (http://www.w3.org/2004/10/swls-workshop-report.html