Reflections on InterPARES
A Pattern Language for Electronic Records

KENNETH HAWKINS

RÉSUMÉ Il y a plus de vingt ans, Luciana Duranti a attiré l'attention sur les caractéristiques intellectuelles et physiques des documents par analogie avec l'architecture. Elle a soutenu que les éléments physiques des immeubles et des documents ont une signification seulement en tandem avec les idées et la culture desquelles ils sont générés. En même temps, les ingénieurs en logiciels, les informaticiens et les architectes de systèmes d'informatique se sont inspirés des concepts de « *Pattern language* » de l'architecte Christopher Alexander afin de faire face au volume grandissant de l'information, à la complexité croissante de celle-ci et aux nouvelles exigences auxquelles ils étaient confrontés. La technologie de l'information comme profession a adopté une approche de recul, en identifiant les patrons (« *patterns* »), en les conceptualisant (« *modelling* ») et en se servant d'outils logiciels afin de déterminer les caractéristiques les plus fondamentales et d'identifier les corrélations entre elles, dans le but ultime de s'en sortir. Les systèmes qu'ils ont conçus ont mené au Web, mais ils ont aussi généré des documents numériques d'une telle quantité et d'une telle complexité qu'ils accablent la profession archivistique.

Kenneth Hawkins examine la seule étude de cas d'InterPARES 2 (2002-2007) qui a tenté de se servir des outils de pointe du Web sémantique – par lesquels les métadonnées ne décrivent pas seulement les documents, mais les rendent utilisables comme programmation informatique – afin de développer un nouveau format de préservation logique pour les documents complexes de l'ingénierie électronique et de la conception. Un bon nombre des patrons conceptuels qui sont bien connus par les diplômés en archivistique qui travaillent avec des documents sur papier – où l'interaction des processus d'affaires et des caractéristiques physiques laissent leur marque indélébile sur les documents et ce, de manière facilement reconnaissable – ont leurs analogies dans le monde des documents numériques complexes. Ceux-ci incluent les standards du Web sémantique et les outils qui offrent des possibilités et qui comportent des risques pour nos lecteurs. La profession archivistique risque de devenir marginale, voire même sans importance, dans l'ère numérique si elle refuse de collaborer avec les professions de la technologie de l'information pour bâtir des systèmes d'archivage en se servant de ces outils.

ABSTRACT Two decades ago Luciana Duranti highlighted the intellectual and physical characteristics of documents with an analogy to architecture. She argued that the physical elements of buildings and documents convey meaning only in

ARCHIVARIA 67 (Spring 2009): 157–188

tandem with the ideas and culture from which they come. At the same time, software engineers, computer scientists, and system architects drew on the concepts of architect Christopher Alexander's pattern language to cope with the increased volume and complexity of information and demands they faced. The information technology (IT) professions adopted an approach of stepping back, identifying patterns, modelling them, and using automated tool support to "thin-slice" only the most salient characteristics and how they correlate in order to manage. The systems they built not only brought us the World Wide Web but also records in volumes and complexity enough to overwhelm the archival profession.

Kenneth Hawkins reviews the one InterPARES 2 (2002-2007) case study that attempted to use cutting-edge tools of the Semantic Web – where metadata not only describes records but makes them actionable within computing environments – to develop a new logical preservation format for complex electronic engineering and design records. Many of the conceptual patterns familiar to students of archival science working with paper records, in which the interplay of discrete business processes and physical characteristics indelibly mark records in instantly recognizable ways, have concrete analogues in the world of complex electronic records. These comprise the Semantic Web standards and tools that pose opportunities and risks to our readers. The archival profession will become incidental or even irrelevant in the digital age if it declines to collaborate with the IT professions to build archival systems using these tools.

"You are not to expect visible proofs in a work of darkness. You are to collect the truth from circumstances, and little collateral facts, which taken singly afford no proof, yet put together, so tally with, and confirm each other, that they are as strong and convincing evidence, as facts that appear in the broad face of the day."[1]

"Software is invisible to most of the world. Although individuals, organizations, and nations rely on a multitude of software-intensive systems every day, most software lives in the interstitial spaces of society, hidden from view except insofar as it does something tangible or useful."[2]

Introduction

The listing and keeping of information that demonstrates the identity of records (specifying, for example, who created them and to what activities and other records they relate) are part of the overall process used to reach a presumption of the records' authenticity. The other critical part of the process

1 Judge Francis Buller to the jury, Donnellan case, England, March 1781, quoted in "Visible Proofs: Forensic Views of the Body," National Library of Medicine, National Institutes of Health, US Health & Human Services, available at http://www.nlm.nih.gov/visibleproofs/index.html (accessed on 26 January 2009).

2 Grady Booch, *Handbook of Software Architecture,* available at http://www.booch.com/architecture/index.jsp (accessed on 26 January 2009).

traditionally falls within the realm of human knowledge. It calls on our ability to evaluate, using logic and reason, the patterns and relationships between characteristic elements of records identity with the events and activities of their original contexts to decide if the records are authentic. These contexts can be of business, culture, technology, or all three. Authenticating records – attesting that they are what they purport to be and that they have not been corrupted – thus involves an implicit but complex mental assessment and correlation of the characteristics of the record or set of records (fonds) by an individual, human agent against accepted or known criteria. This has been the role of the jurist, historian, archivist, and records manager for centuries but will soon be impossible for any of them to perform in ways that are familiar.

Ad hoc authentication of many objects, including traditional artifacts and records, has often been done effectively and accurately by a single expert in a blink. Through a nearly unconscious process of cognition, dubbed "thin-slicing" by psychologists, the full range of information necessary to make a decision is filtered down to a few critical elements and processed instantly; experts may tell at a glance, with as much accuracy had all the information relevant to judging been evaluated over a longer period. They know whether the elements of the object, taken individually and in relation to one another, "add up," "look right," or just the opposite. For example, the J. Paul Getty Museum acquired, at considerable expense, what purported to be an ancient Greek *kouros* (or sculpture) of a young male nude. Experts from several fields examined it from every angle and pronounced it authentic. Others took only one look at it and reacted instantly with impressions across a range of negation: from the ancient work looked "fresh" to feelings of "intuitive repulsion." Hunches gave way to certitude when further analysis showed the *kouros* to exhibit signs of work from different styles and time periods. It was learned the patina it carried could be duplicated in weeks with potato mold. Letters related to the provenance of the work proved to have flaws within their own provenance: one letter dated 1952 exhibited a postal code not issued until two decades later. The Getty's catalogue entry now reads: "About 530 BC, or modern forgery." In 1983 at least three prominent World War II historians endorsed the authenticity of a series of small manuscripts purported to be the personal diaries of Adolph Hitler. When different experts examined the collection they quickly deducted its fraudulent origin, simply by noticing that it was printed on modern paper and that the letter "F" was used instead of "A" in a monogram set in old typeface.[3]

In both cases instantaneous assessments using limited but critical or

3 Malcom Gladwell, *Blink: The Power of Thinking Without Thinking* (New York, 2005), pp. 3-8; Robert D. McFadden, "Skepticism Growing Over 'Hitler Diaries'," *The New York Times* (25 April 1983), p. 1.

telling inputs and criteria, yielded results as good as, or better than, exhaustive examinations - and did so faster. The human mind excels at such rapid critiques, and examples abound where the ability to do a quick read of an object or situation, decide, and act, mean the difference between truth or fraud, success or loss, even life or death.

But the same is becoming less true for those who keep records. Their methods used to authenticate records – whether for business or archival purposes – indeed the very basis of their professions, mean that they face issues that challenge the future and very survival of their vocation. What has been termed the most important industry in the world, software engineering, now touches every aspect of life, commerce and culture,[4] and at each touch digital records are created. The types and complexity of digital records multiply constantly and their volume increases exponentially, even as it is left to human ingenuity for the most relevant of these records to be found, understood, and used. But many of the characteristics of records that people relied upon to authenticate records are now latent in digital objects, that is, not human readable. Multiplying threats, tampering, and attacks against records are also concealed until it is too late. One significant example of concern is that voiced by critics of the so-called direct recording electronic voting systems now in use across the United States. The problem, writes a Stanford University computer scientist, is that "paperless e-voting technology is almost totally opaque."

> No one can scrutinize critical processes of the election, such as the collection of ballots and counting of votes, because those processes occur invisibly in electronic circuits. Voters have no means to confirm that the machines have recorded their votes correctly, nor will they have any assurance that their votes won't be changed later.[5]

To the archivist lacking expertise in software engineering, computer science, or information technology (IT), the most telling elements of records may as well be invisible. Traditional means of authenticating records, whether done *ad hoc* by a single archivist or in studied deliberation by a whole staff of archivists, are inadequate to the challenge. Where once the characteristic elements and patterns of records were explicit and the reasoning process to authenticate them was latent, now the characteristic elements and patterns of records are latent and the logical reasoning process to authenticate them must be made explicit. Unless the archival profession both radically shifts how it sees these changes and addresses this work, it will be rendered irrelevant or incidental in the digital age.

4 Booch.

5 Testimony of David L. Dill before the Commission on Federal Election Reform (Carter-Baker Commission), 18 April 2005, Hearing, American University, Washington, DC.

While the archivist is no longer able to do the heavy lifting of records authentication using traditional means, concerned technologists have developed a myriad of "solutions" focused on media-checks, bit-counts, encryption, or controlled access. These approaches, while valuable, are not sufficient.[6] While they may help establish the integrity of a record (that it remained free from tampering), they cannot corroborate its identity. They do not have the analytical value of an approach founded on logic, picking up from the approach used by archivists who correlate the salient characteristics of a record in a blink. Authentication is a complex, not monothetic, argument; it must allow for the consideration of multiple variables. This is not to say technical and computing approaches are without value. Indeed, given the origins of digital records within computing environments and their volume, complexity, and suitability to machine processing, an approach based in computer science and complex archival cognition is both necessary and promising. But whereas the IT professions have developed and standardized processes and tools with which to manage the volume and complexity of their information needs and demands, archivists have barely joined the conversation in which their insights and concepts might advance similar means of creating and preserving authentic electronic records.

This article examines the promise that the so-called semantic technologies hold for the authentication and preservation of electronic records, and suggests preliminary requirements to help move from promise to reality. It begins in the first section with the pattern language concepts that find and validate (i.e., authenticate)[7] desirable designs in object-oriented computer software programming, system modeling, and Web architecture. The section includes analogies to records keeping and shows how the idea that documentary artifacts reflect the context of their creation resonates with technical leaders. These pattern language concepts drive the development of today's Semantic Web, where the knowledge contained in the fixed information assets (that is, *records*) held by domains (from single business firms to the World Wide Web) is being made available for discovery by, reasoning over by, and use of machine agents acting on behalf of people. It is this approach to commerce, communication, and culture that now generates the mass of records with which archivists have to contend. More importantly this section shows how

6 International Research on Permanent Authentic Records in Electronic Systems [hereafter InterPARES], *Authenticity Task Force Report*, in *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project* (Vancouver, British Columbia, 2002) [hereafter *ATF Report*], p. 2. This report is available at http://www.interpares.org/book/index.cfm (accessed on 26 January 2009).

7 While the IT professions do not use the specific term "authenticate" or its variants, there are common themes between their use of pattern language concepts to solve problems, and usage of the term in archival science and diplomatics. This paper aims to show where this union can help address the challenges facing those who keep electronic archives.

professional communities, first within architecture and then IT, successfully contended with the complexity and volume of information, and needs that threatened to overwhelm them. Each group stepped back, identified commonly occurring but discrete patterns, modelled them, and used automated tool support to "thin-slice" only the most salient characteristics. When it comes to authenticating and preserving complex and voluminous digital records, the capabilities in use by the IT professions to build the Semantic Web offer a way forward to today's archivists.

In the second section we examine how assuring the preservation of authentic electronic records challenges one domain out of many: that of high-tolerance, high-assurance, science-based manufacturing across government, industry, and academia, which relies on millions of complex digital engineering records.[8] A software engineering experiment conducted by InterPARES 2 researchers, Case Study 19, "Preservation and Authentication of Electronic Engineering and Manufacturing Records," is unique within the literature in that it examines the problems and opportunities where archival theory, diplomatics, and the Semantic Web intersect. It tested the use of open source formats, semantically coded metadata, and reasoning software for the preservation and automated authentication of such records. Its message-based approach provided glimpses of authentication considerably more effective than media, bit-count, or the *ad hoc* checking of discrete provenance attributes by a single person.

We then ask whether the findings of the experiment in complex engineering records can be extended by using the pattern language of documentary forms as a basis for assessing the authenticity of less complex records. We also consider whether this message-based approach can be supported by the automatic generation or encoding of metadata attributes at record creation and thereafter for processing by relatively simple and efficient reasoning programs.

Solutions to technical problems start with dialogue. The computer science, enterprise architecture, and software engineering industries are well along in devising capabilities for other communities that archivists could use. This article aims to broaden the awareness of common concerns and solutions beyond the few individuals and groups that see the connections. It envisions the development of a patterns-based authentication and preservation method, using simple logic and semantics to query reliably populated attributes carried with electronic records through their life cycle. The outcome could be simple, scalable automated checks, based on business knowledge expressed in rules, to validate and preserve authentic electronic records.

8 Such records include those produced from computer-assisted design (CAD), computer assisted engineering (CAE), and computer assisted manufacturing (CAM).

Pattern Language Concepts

"All acts of building are governed by a pattern language of some sort, and the patterns in the world are there, entirely because they are created by the pattern languages which people use."[9]

"The form of a document is of course both physical and intellectual. An analogy with architecture may help clarify this vital concept."[10]

Over the last quarter century one of the most significant and generative trends in the information technology (IT) industry, across the domains of software engineering, enterprise architecture, and representation or markup languages, is the influence of a paradigm for identifying and establishing quality in built spaces. Out of the welter of competing approaches for designing and building software to meet user needs, computer specialists came to appreciate the simple yet elegant ideas of a Professor of Architecture at the University of California, Berkeley, named Christopher Alexander. Alexander writes about what makes a built space *alive*, whether it is a house that makes one feel at home, an office building that people enjoy working in, or a small sun-filled terrace that nourishes the spirit as it offers a pause from daily activities. Place by place, he called out not just the discrete elements that characterize built spaces that are alive but especially the pattern(s) of relationships between and among those elements, *and* the repeating patterns of human events to which they give rise. To IT practitioners interested in building and delivering software products that were efficient, reusable from project to project, scalable in meeting needs of differing sizes and volumes, and in discovering and establishing repeatable methods for doing this, Alexander's writings rang true. The IT systems now being built by practitioners generate electronic records of such volume and complexity to overwhelm the archival profession. Archivists must learn how architects and then engineers within IT used Alexander's concepts in order to contend with the volume and complexity of problems that began confronting them two decades ago.

Among the examples Alexander provides to illustrate the language that guides patterns of relationships between discrete physical elements or characteristics and the repeating events or behaviour of people in them, is the medieval Gothic church. We know it as a Gothic church because of the characteristic layout and dimensions of its elements: a long nave flanked by aisles running parallel to, and narrower than, the nave, a transept crossing them at right angles near the front, lines of evenly spaced columns separating nave and aisles, and supporting buttresses that lead upward to vaults concave

9 Christopher Alexander, *The Timeless Way of Building* (New York, 1979), p. xi.
10 Luciana Duranti, *Diplomatics: New Uses for an Old Science* (Metuchen, 1998), p. 41.

in space. Higher up large stained-glass windows line each wall of the nave and admit multi-coloured beams of light into the sanctuary. Though there is some variance, these patterns repeat from one Gothic church to another, and the physical elements also accommodate repeating patterns of events within the space with similarly unique and distinguishing characteristics. Likewise, a small terrace set adjacent to a country house on its southern exposure, enclosed by low walls comfortable to sit upon, its stone pavers warming to the winter sun, an espaliered fruit tree trained against its wall, presents patterns both between the discrete physical elements that say "sun-warmed terrace," and between these patterns and predictable human rituals of respite, relaxation, and pause. The pattern of elements does not "cause" the pattern of events or vice versa; "the total pattern," elements and events together, is an expression of culture. "It is invented by culture, transmitted by culture, and merely anchored in space."[11] Minor variations in the elements that undergird the patterns matter less in establishing the place than the repetition of long-lived patterns of events and behaviours. But in either church or terrace, take away any set of these patterns *without replacing it* and the quality of the place becomes something different, something less true to itself. The essential, authentic character of any given place is the dialogue between physical elements and behavioural events: "The character of a place, then, is given to it by the episodes which happen there."[12]

Having established these concepts Alexander provided a workbook of 253 concrete examples from which architects could work to create humane, livable homes, workplaces, built landscapes, and towns.[13] Each design "pattern" was presented as a short case study or record of design solutions for common problems, balancing the interplay of physical elements and social relationships to resolve conflict between them. Each pattern thus became a tool that could be used every time the same basic requirement showed up. Alexander's intent was to disseminate his findings as the basis for a new way of designing and building. With it, anyone could render a built environment at once true to its character, economical, alive, and even beautiful. Among architects and planners his ideas are embraced by proponents of the New Urbanism and Smart Growth movements, which emphasize mixed land use for work, home, and play, walkable communities, sustainable economies, etc.; they are not, however, regarded well by the mainstream architecture establishment.[14]

11 Alexander, *The Timeless Way of Building*, p. 92.

12 Ibid.

13 Christopher Alexander, *A Pattern Language: Towns – Buildings – Construction* (New York, 1977).

14 Wikipedia, *s.v.*, New Urbanism, http://www.wikipedia.org (accessed on 26 January 2009); Andres Duany, Elizabeth Plater-Zyberk, and Jeff Speck, *Suburban Nation: The Rise of*

Alexander's ideas have found a receptive audience among software engineers and architects. Design patterns make sense for business reasons because they allow for the efficient production and implementation of software, based on best practices and proven solutions. The first to explicitly embrace pattern language concepts was a group of software designers working in object-oriented programming in the 1980s and 1990s. Object-oriented (OO) design consolidates data (expressed as attributes) and behaviour (expressed as methods) into discrete entities called objects. An object for person in a computer system might have the attributes *Name / SocialSecurityNumber / DateOfBirth / Gender*. Each attribute then has (its own or access to) a corresponding method to change or set the attribute's value, and to get or communicate the result. By encapsulating data and behaviour together in objects, OO design and programming assembles building blocks of functionality that may work independently of, or in accordance with, one another, by means of messages, without having to know or care how any one object does its own part. Clear interfaces between objects also allow for the separation of types or implementations of functionality, such as the user interface, business or processing logic, and data. Such rigorous specification also allows for the calling out of points where records are created, set aside, and used to support business processes. The object that stores the personal information in the example above serves both as a record of it and a reference point for other processes, such as the crediting of payments or verification of identity. The OO approach is thus well-suited to real business environments, where computing systems that interoperate are distributed from each other, such as the World Wide Web, which was largely built using OO technology.[15]

To keep track of the best approaches to meet computing needs in that fast-evolving environment, software engineers increasingly adopted the concepts and tools of a pattern language. Apart from its practical "tool-kit" aspects, the pattern language approach as carried forward by software designers and system architects, also embraces the quest for ascertaining and documenting quality or authenticity in each particular "space" it considers. Alexander spent considerable effort discussing how the union of characteristic elements and repeating patterns of events distill into the essence of a place (or a thing), giving it "the quality without a name" – a quality that is by turns alive, desirable, beautiful, authentic, and even true.

In the world of living things, every system can be more real or less real, more true to

Sprawl and the Decline of the American Dream (San Francisco, 2001), pp. 183-214, 268, 273.

15 Matt Weisfield, *The Object-Oriented Thought Process*, 2nd ed. (Indianapolis, 2004), pp. 5-14*ff.*

itself or less true to itself. It cannot become more true to itself by copying any externally imposed criterion of what it ought to be. But it is possible to define a process which will tell you how the system can become more true to itself, in short what it "ought to be," only according to what it is.[16]

On the face of it the quality that is desirable to the software engineering community might relate to something gained from the codification of best practices, such as interoperability or what one engineer terms "the holy grail of software development: software reuse."[17] It might relate to a software design that captures the so-called "-ilities": reliability, maintainability, adaptability, scalability, etc. Beyond these, how could an enterprise architecture model or software release be deemed beautiful, true to itself, or to possess "the quality without a name?"

The key lies in the appreciation the OO community's leaders have for the semantic underpinnings of Alexander's arguments, and for those evident in the objects and processes of their own disciplines. Not only have they adopted architectural similes, they also defined a process just as Alexander envisioned. They created repeatable methods of identifying and duplicating quality in systems across a spectrum of computer engineering contexts by explicitly documenting the characteristic relationships between, and behaviour of, otherwise discrete elements, entities, objects, services, the classes they comprise, and the constraints and cardinalities that guide them. The semantic in OO design patterns is found not only in the engineering or business domain-specific "place" that they document, tracing the circumstances it presents to the pieces that go into solving it and how they relate to one another, but it is also in the ability to specify, using increasingly standardized notations, the relationships between all kinds of classes and the characteristic logical patterns that hold them together in a way that allows the design, testing, and building of information systems. "Indeed," notes Grady Booch, Chief Scientist at IBM's Rational Software and co-author of the *Unified Modeling Language*, "every well-structured software-intensive system is full of patterns, ranging from idioms that shape the use of a particular programming language to mechanisms that define the collaboration among societies of objects, components, and other parts."[18] And as archivists know (or ignore at their peril), when information systems begin creating records, the reach of patterns involved encompasses not only the technology at work but the business processes and agents technology serves, and how patterns imprint

16 Alexander, *The Timeless Way of Building*, p. 28.

17 Weisfield, p. 237.

18 Booch; see also by Booch, "On Architecture: From Small to Gargantuan," available at http://www.informit.com/articles/article.aspx?p=517211 (5 July 2006) (accessed on 26 Ja nuary 2009), originally published in *IEEE Software Magazine* (available at http://www.computer.org/software).

themselves on the records produced. Grasping the relationships between these sets of patterns is the first step toward using them to create and preserve authentic electronic records.

Simultaneous with the growing influence of pattern language concepts and instruments in software and enterprise architecture during the 1990s, leading proponents of OO software design and engineering recognized the need for a way to graphically model the concepts, behaviours, and relationships of the core entities in any given business domain, independent of the details of hardware and software capabilities. Unlike specific programming languages, development techniques, and implementations crowding the IT marketplace, the modelling language had to put the characteristics unique to the business space at the forefront, while at the same time abstracting them into a notation or form of expression that was normalized, repeatable, and could be used by any number of engineering approaches to propose solutions. Only then could a documented path be drawn and maintained from the essential characteristics of the business and its requirements to the specific IT functions capable of servicing them, using this "thin-slice" to filter out the extraneous elements and information that could otherwise overwhelm comprehension.

The *Unified Modeling Language* (UML) emerged as the premier modelling tool for designing, describing, and documenting OO software systems. In keeping with the need for a language independent of specific software platforms, the UML was first adopted as an open standard in 1997 by the consortia of companies that support interoperability between OO systems, the Object Management Group (OMG), and received a major update in 2004. The UML consolidated or replaced a dozen competing OO modelling methodologies and is supported by all major software modelling and some software coding tools. Widely used by software engineers across industry and government, the specification maintained by the OMG defines the UML as "a graphical language for visualizing, specifying, constructing, and documenting the artifacts of distributed object systems."[19] Artifacts are the UML's representations or notations for objects and the classes they can comprise, the attributes and behaviours they exhibit, their relationships to one another and to their context(s), and the constraints or rules that condition their behaviour and relationships.[20] Each feature in the UML is presented graphically in a notation reserved for it exclusively. A well-done UML model captures only the artifacts most important to the business and its requirements, and

19 Object Management Group, *Unified Modeling Language* (UML), version 2.0, available at http://www.omg.org/technology/documents/modeling_spec_catalog.htm#UML (accessed on 26 January 2009). The UML is also available from ISO as ISO/IEC 19501.

20 Martin Fowler, *UML Distilled: A Brief Guide to the Standard Object Modeling Language*, 3rd ed. (Boston, 2004), pp. 1-9, 35-52.

provides the blueprint for the construction, testing, and maintenance of an IT system positioned to address those requirements. "UML models," notes one proponent, "are situated culturally and socially in the organizations and processes that they both reflect and shape."[21]

While the UML has the capability to model a variety of dynamic characteristics and behaviours for OO entities pertaining to any business process, the core entities binding a software system to its business context are modelled in static views by class diagrams.[22] The shared features of a class provide evidence of its identity and responsibilities or roles, capturing the vocabulary of the business space. The features documenting the relationship(s) between classes capture the grammar of the business space. Many relationships between classes signify dependencies and constraints that are derived from business requirements. Taken together the web of relationships and classes begins to form a topography of meaning for the business domain they model; archivists examining it later can use this evidence to help preserve authentic electronic records.

Pattern Language Concepts and the World Wide Web

Perhaps the most intriguing and complex diffusion of pattern language concepts has occurred where its impact could be greatest: the evolution of the World Wide Web. As proponents of OO programming, enterprise architecture, and model-driven development learned how to detect and make explicit the patterns at work in graceful solutions, those responsible for the world's first global web of hypertext found they had to do the same. The phenomenal growth in the number of web pages available on the public web drove home the conclusion that such a body of latent knowledge, presented foremost for humans to read, impeded comprehension and usefulness. "We are drowning in a sea of data which occasionally is generously referred to as 'information'," read the call for a meeting on the issue in 2005. "But the truth is that almost all of it must be interpreted by humans to be of any use."[23] Today's Web is glutted with documents and media for humans to discover

21 Paul Evitts, *A UML Pattern Language* (Indianapolis, 2000), p. 207.

22 At their semantic core, class diagrams specify the relationship between the elements that make up a given business space and the ground rules for relations, interactions, and dependencies. Because of their normative role in the development and operation of OO systems, many so-called design artifacts (such as class diagrams and data models) function as documentary entities, that is, they have fixed form and content. Although the setting aside and keeping of such documentary evidence is considered best practice amongst OO proponents, archivists have given little consideration to their usefulness for archival purposes beyond the day-to-day management of business information.

23 Semantic Technology Conference, 6-9 March 2005, San Jose, CA. Proceedings available at http://www.semantic-conference.com/default.html (accessed on 21 April 2009).

and consume on an *ad hoc* basis. Given that the Web hides much of what is required for humans to discover, examine, and reach conclusions about information, its audience is largely left without the tools to turn information into effective knowledge except on a piecemeal basis – not unlike the challenges faced by the archival profession.

> The World Wide Web as it is currently constituted resembles a poorly mapped geography. Our insight into the documents and capabilities available are based on keyword searches, abetted by clever use of document connectivity and usage patterns. The sheer mass of this data is unmanageable without powerful tool support.[24]

The approach to develop such tools underway by the World Wide Web Consortium (W3C), the keeper of the key open standards driving the Web, its partners in other standards bodies (including the ISO and OMG), industry, government, and academia, is to evolve a Semantic Web from the current one. The vision is a globally integrated network of self-describing, interoperable data unlocked from relational databases, XML documents, spreadsheets, content management systems, etc., and made available for "re-purposing" by computers and humans. The Semantic Web is to be a "navigable space," where named associations link data and information given "well-defined meaning" by humans but that are encoded into standard formats that computers can discover, integrate, reason over, and place continuously into service.[25] In other words, the Semantic Web can make the kinds of patterns and correlations that humans make to ascertain and understand small volumes of information automatically effective with much larger volumes of information.

The Semantic Web

The Semantic Web vision is holistic but from the start its realization has been from the bottom up. The architects behind it are keenly aware that individuals cannot encode legacy and day-forward assets page by page or object by object. The Semantic Web's progress has relied on adoption of its tools by specific domains and is being built from a tiered architecture, based on open standards, of data representation with ascending layers of meaning. In 2000, the architecture of the Semantic Web looked like this:

24 World Wide Web Consortium [W3C], *OWL Web Ontology Language Guide*, *W3C Recommendation*, Michael K. Smith, Chris Welty, and Deborah L. McGuinness, eds. (10 February 2004), available at http://www.w3.org/TR/owl-guide/ (accessed on 26 January 2009).

25 Tim Berners-Lee, "Semantic Web," keynote, XML 2000 (6 December 2000), Washington, DC, available at http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide1-0.html; Tim Berners-Lee, James Hendler, and Ora Lassila, "The Semantic Web," *Scientific American* (May 2001), pp. 34-43.

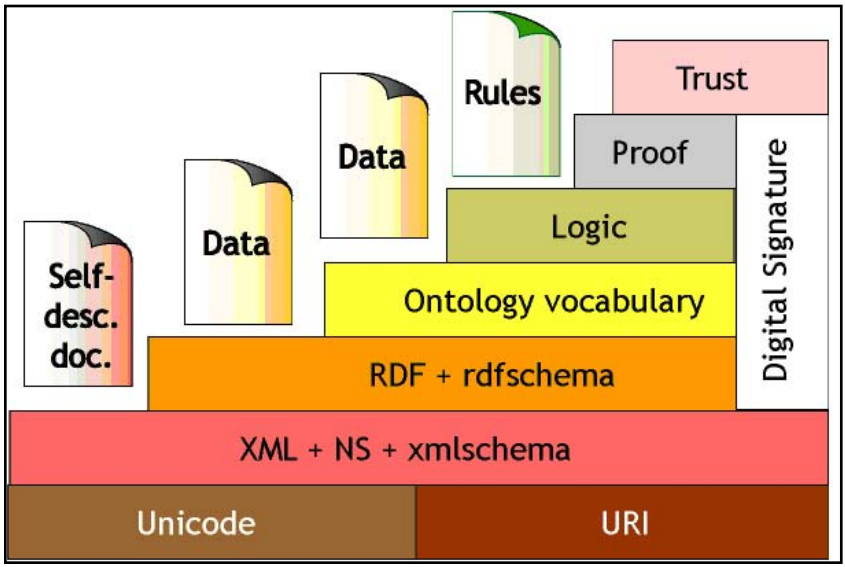

Figure 1: The Semantic Web[26]

Fully realized, the Semantic Web architecture is designed to support the representation of data and information from the atomic level of its character sets and locations to the meaning(s) inherent in how classes relate to each other, the rules that guide the relations, the proofs that can be derived, and the trust that can be assured. In other words, it is to follow a pattern much like Alexander's in which the specification of the physical elements and structures is followed by how they relate to one another to inform the character of a given space or domain. As it stands, some communities of practice have made progress in realizing this pattern, while others have barely recognized its importance or have remained stationary at its lower levels.

As noted, development of the Semantic Web has been incremental. At the turn of the century the foundation tiers of the architecture in Figure 1 encompassed the non-normative elements of the World Wide Web; standards were adopted and implemented for:

- a common character encoding scheme (Unicode);
- persistent identifiers fixing the location of resources (URIs);
- the syntactic representation of document structure or composition (XML).

26 Deborah McGuinness and Mike Dean, "Substance of the Semantic Web," Semantic Web Applications for National Security (7 April 2005), Arlington, VA, available at http://www.daml.org/meetings/2005/04/pi/Substance.pdf (accessed on 30 March 2009).

XML, derived as it is from an encoding schema originally intended to prepare documents for publication,[27] provides for validation of well-formed syntax through interactions between Document Type Definitions (DTDs) and documents, but does not show semantic relationships between any given set of elements within a document or between sets of documents. XML cannot express shared meanings of terms or concepts used to describe objects or resources on its own, nor can it convey the meaning implicit in the associations between them. However, the syntax of more semantically robust metadata schemas conveying these relationships can be encoded and transported (serialized) in XML. The *Resource Description Framework* (RDF), formally recommended by the W3C in 1999 (and since extended and revised), allows the creation of metadata to express the properties, property values, and class memberships of Web-accessible objects or resources. The "resource" being described could be a web page, a document, an image file, a data attribute or a row in a database, essentially any real world thing represented by data or information. More importantly RDF also specifies the relationships a resource has to other resources using a subject/predicate/object "triples" notation *that is machine-processable*. The real world relationships between a person's credit score, as well as the date range of their job tenure or residency, can and is stored as a record and processed for business purposes using RDF metadata.[28] Each element within this pattern is considered a resource and is distinguished by its own URI, as are the relationships linking them. Mapping the "navigable space" of resources, associations, and patterns using the language and syntax of RDF and its underlying components, makes it ready for a variety of semantic work (integration, translation, reasoning, inference, proof) that computers implementing the higher tiers of the Semantic Web architecture currently do.[29]

Just as an XML-DTD can be joined with the document it constrains or be available to it across a network space, RDF resources do not by necessity have to be physically contiguous to function. In fact the distributed nature of the Web means that usually they will only be logically contiguous. As RDF

27 Robin Cover, "The SGML/XML Aversion to Semantics," *Cover Pages* (27 September 2000); available at http://xml.coverpages.org/sgmlEschewsSemantics.html (accessed on 26 January 2009).

28 One of the widest implementations of RDF is the *Extensible Metadata Platform* (XMP) in consumer and professional digital photography, which memorializes and makes available for subsequent use, a variety of metadata in a "sidecar" embedded in .jpg and .raw image files generated by digital cameras and other imaging devices and platforms. See Wikipedia, *s.v.*, Extensible Metadata Platform (http://www.wikipedia.org).

29 McGuinness and Dean; Jim Hendler, "From Atoms to Owl's: The New Ecology of the WWW," keynote, XML 2005 (15 November 2005), Atlanta, GA, available at http://www.cs.umd.edu/~hendler/presentations/XML2005Keynote.pdf (accessed on 26 January 2009).

has advanced, not only has its syntax been simplified but specifications to control vocabularies (RDF schemas), the export of discrete URIs as first class objects to purpose-built databases (or "triple stores"), and RDF provenance, have advanced or been adopted as standards.[30] The linking of semantically related records to support the operation of otherwise separated systems is a trend that is not going away. Is the archival profession prepared to contend with the truly radical types and volumes of records this trend is bringing?

In the last several years, the W3C and a number of business domains have made progress in building the middle tiers of the Semantic Web[31] that allow the expression of meaning inherent in the relationships between objects (schemas and URIs) and resources (records). The proliferation of URIs specifying the identity and relationships of discrete resources, in turn calls for a separate framework that allows the formal definition of vocabularies for objects or resources, their classes, the relationships linking them, as well as the use of different terms for like concepts. Borrowing a word from philosophy, the architects of the Semantic Web chose the name "ontology" for the tier addressing this and have begun to implement it.

Web Ontology Language (OWL)

Making the meaning behind objects and resources operational using ontologies has been at once a top-down and a bottom-up process. In 2004 the W3C recommended the *Web Ontology Language* (OWL), noting that it adds to RDF's capabilities for describing objects or resources, their properties, and associations. OWL is a declarative language, based on description logic, that can express specialization/generalization hierarchies, relationship cardinalities (including the many-to-many kind expected in a web), and also has the ability to restrict ranges of values and specify where properties are equivalent or unique. If this sounds like UML, there is a basic difference. UML models these patterns; OWL encodes their description using characters and embedded logic that computers can process in tandem with actual information objects and resources. These capabilities provide needed support for indexing and searching during the active use of resources; less attention has been given to how they may persist and service other requirements following this phase, including the ongoing requirement to manage complexity and

30 Ibid.; Jeremy J. Carroll, Christian Bizer, Pat Hayes, and Patrick Stickler, "Named Graphs, Provenance and Trust," Semantic Web Foundations session, Fourteenth International World Wide Web Conference, 10-14 May 2005, Chiba, Japan; available at http://www.2005.org/cdrom/docs/p613.pdf (accessed on 21 April 2009).

31 Nigel Shadbolt, Wendy Hall, and Tim Berners-Lee, "The Semantic Web Revisited," *IEEE Intelligent Systems* (May-June 2006), p. 98, available at http://www.consortiuminfo.org/bulletins/semanticweb/php (accessed on 6 Janurary 2009).

volume of archival resources.[32] Addressing the broad requirement to enable better comprehension of any large body of information, the specification states: "OWL is intended to be used when the information contained in documents needs to be processed by applications, as opposed to situations where the content only needs to be presented to humans."[33] This statement is key because it addresses the crux of the challenge: how can humans, once able to assess and comprehend a record based on small volumes and explicit physical characteristics, be able to contend with infinitely greater volumes of records whose essential characteristics are both highly complex and largely invisible?

Heralded as "an ontology language for the Web," OWL has been a domain-driven undertaking in actual conception and execution. The standards creation and adoption process at W3C relies on experts from domains (areas of business) that require common standards to advance real world business requirements. The individuals contributing to, and implementing, OWL have come from academic and commercial organizations (including life sciences, pharmaceuticals, manufacturing, computer science), government agencies (civilian and defense), and other standards bodies (OASIS, ISO, OMG). Discussions of what is meant by ontologies and why they are needed centre on their origins in, and usefulness to, specific domains. "An ontology defines a common vocabulary for researchers who need to share information in a domain," write Natalya F. Noy and Deborah L. McGuinness. "It includes machine-interpretable definitions of basic concepts in the domain and relations among them." Far from attempting to create a top-down ontology of everything ("clearly impossible," admits Tim Berners-Lee), the OWL specification and the larger undertaking rely on the creation of ontologies useful to discrete domains. When placed within a distributed environment, such ontologies can help integrate stores of data across the domain and need only make broader linkages where it makes sense. Thus again the way forward to develop IT systems capable of managing otherwise overwhelming complexity and volume has followed an Alexandrian approach of keeping to the patterns characteristic of "places" well-known by subject matter experts. Pointing to the significant and successful development of ontologies in the life sciences and other domains, Berners-Lee and his colleagues note that, "[t]he ontologies that will furnish the semantics for the Semantic Web must be developed,

32 One exception, published since this paper was drafted, and focused on collections management, is the work to implement a semantic metadata repository for information retrieval at the National Archives of Korea. See Tony Lee, Jin Woo Kin, Bok Ju Lee, Kyu Hyup Kim, and Yoon Jung Kang, "Use Case: Semantic MDR and IR for National Archives," available at http://www.w3.org/2001/sw/sweo/public/UseCases/SaltLux-NAK/ (accessed on 8 September 2008).

33 W3C, OWL Web Ontology Language Overview, W3C Recommendation, Deborah L. McGuinness, Frank van Harmelen, eds. (10 February 2004).

managed, and endorsed by practice communities."[34] Any archivist who has managed records from a specific organization or business will appreciate the value of making the concepts, terms, and values that populate its provenance available to the operational tools of its information technology. Leveraging these resources and tools for archival work is the next logical step.

With the middle tiers of the Semantic Web architecture solidly in place (and named by specification in later, more complex versions of the model), the upper tiers related to expressing rules, providing proofs, and attaining trust are now coming into view.[35] Even if the vision of "one huge database"[36] of globally interoperable data supporting such higher functions is never achieved, to date the work in domains has shown how IT capabilities can be driven forward in beneficial ways where pattern language concepts are present. The United States Air Force has developed an OWL ontology for "situational awareness," to know about the identities and characteristics of objects in a domain, coupled with a rules engine to know with certainty the real-time relations of objects in the same domain. Making these patterns explicit derives from business requirements with serious implications:

> For example, simply knowing that there is a west-bound airline and an east-bound airline on the radar screen is not as important as knowing that the two planes are "dangerously close" to one another. In this case, "dangerously close" is a relation between two objects that must be derived from sensory data, although the data by itself says nothing about the concepts of "closeness" or "dangerous."[37]

At the National Cancer Institute, OWL ontologies provide semantic maps of the kinds of cancers occurring in various anatomical sites when a

34 Kendall Grant Clark, "The Semantic Web is Closer Than You Think," *XML.Com* (20 August 2003), available at http://www.xml.com/pub/a/2003/08/20/deviant.html (accessed on 26 January 2009); Natalya F. Noy and Deborah L. McGuinness, "Ontology Development 101: A Guide to Creating your First Ontology," 2002, Knowledge Systems Laboratory, Stanford University, Stanford, CA, available at www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html (accessed on 26 January 2009); Andrew Updegrove, "The Semantic Web: An Interview with Tim Berners-Lee," *Consortium Standards Bulletin* (June 2005); Shadbolt, Hall, and Berners-Lee, p. 99.

35 Besides RDF and OWL, these specifications include RDFS (schema), RIF (rules), and SPARQL (queries). Shadbolt, Hall, and Berners-Lee, pp. 98-101; Elisa Kendall and Evan Wallace, "The Ontology Definition Metamodel: A Tutorial," OMG Technical Meeting, 27 September 2006, Anaheim, CA.

36 Tim Berners-Lee, quoted in Updegrove; W3C, "W3C Semantic Web Activity," available at http://www.w3.org/2001/sw/ (accessed on 11 February 2009).

37 Christopher J. Matheus, Mitch M. Kokar, Kenneth Baclawski, and Jerzy Letkowski, "Constructing RuleML-Based Domain Theories on top of OWL Ontologies," Versatile Information Systems, Inc., Northeastern University, Western New England College, 2003, n.p., available at http://www1.coe.neu.edu/~kokar/publications/RuleML03.pdf (accessed on 26 January 2009).

given pattern of conditions is present. Each object, class, and association representing real things and their correlations in this setting is modelled in UML and fits into the overall enterprise architecture that bridges directly to the agency's mission to relieve suffering from cancer.[38] In both systems fixed documentary entities that an archivist would recognize as records, participate in a complex dynamic of services and transactions that proceed from, and actuate, both social and technical patterns. The qualities that pattern languages can support clearly are not divorced from those of the real world.

The above are only two cases of many where the "Web" by definition is not worldwide but domain-specific and nonetheless extraordinarily complex. The space in which patterns are being detected, made explicit, and reasoned over may even be limited to a single function or activity, captured in a single RDF/OWL expression. This approach may be what makes the intractable Web of data and information tractable. Based on the human mind's ability to "thin-slice" its way through a tangle of data, abstract the elements most vital to a finding or decision, the pattern language, object-oriented approach to software architecture, made operational with modelling and encoding tools designed for semantic expression, may also enable people to preserve authentic electronic records forever. The next section examines the significance of an engineering experiment that tested Semantic Web tools to do just that with the complex digital records generated by modern computer-aided design and computer-aided manufacturing systems.

Archives and the Semantic Web

"The Semantic Web is a Web of actionable information."[39]

"Data with attributes do not constitute a knowledge form."[40]

Why should archivists look to the pattern movement in Information Technology? First, because it is in IT that digital records are being created amidst a semantic web of models, schemas, ontologies, and business rules. The wide use of object-oriented systems and semantic technologies in the arts, science, industry, and government domains clearly shows that information objects are treated as entities requiring management across continuums including moral, juridical, and chronological. It is apparent that these domains are creating and maintaining, if not always preserving, digital objects whose characterization encompasses, but often extends considerably,

38 See "caCORE Overview,"http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview (accessed on 21 April 2009).

39 Shadbolt, Hall, and Berners-Lee, "The Semantic Web Revisited," p. 96.

40 InterPARES 2 Researcher, Case Study 19, *Final Report*, 2005, p. 9.

the discrete identity and integrity attributes traditionally associated with records. While these digital objects may not rise to the level of a canonical record envisioned by general diplomatics, in essential terms they function as documentary entities having fixed form, stable content, an identifiable context, being set aside for reference and to enable subsequent actions.[41] More importantly, without developing the competencies to preserve authentic electronic records, the archival profession will languish.

The pattern language concepts that have informed software architecture, especially object-oriented (OO) software engineering, model-driven development, and semantic web technologies, resonate with the concepts of archival science and diplomatics. Traditional approaches to archival concepts and functions have always taken advantage of, indeed relied upon, the value imparted to records by the business processes and technologies of the creating environment: it is no different today.[42] How documentary entities relate to one another (e.g., a record to a register, both record and register to a fonds), to the classes they are part of, to system entities and services, and their operations and behaviours, the subsequent action they support, are all characteristics that can and are being modelled, documented, and operationalized within a Semantic Web using tools like *Unified Modeling Language* (UML) and next-generation XMLs such as the *Resource Description Framework* (RDF) and *Web Ontology Language* (OWL). The fact that business transaction activities and system interactions of documentary entities cohere into repeating patterns is significant when it comes to records creation and use. It also means that the patterns modelled and used in system design and operation can support archival needs. Because these patterns imprint evidence of characteristic relationships between object classes on the documentary entities created by the system (relationships commonly overlooked or left implicit by archival metadata) and themselves are fixed and set aside as entities to manage them during system operations, they can provide semantic technologies with what they need to reason over these resources for discovery, authentication, use, and preservation of electronic records.

Given the volume, complexity, and stranglehold of proprietary software on burgeoning digital assets/records, it is risky to assume that archivists, without a radical shift in how they approach and carry out their work, will continue to be the primary agents to conduct archival functions, do so on an *ad hoc* basis, be qualified to assess domain-specific metadata designed to be processed by

41 Luciana Duranti and Kenneth Thibodeau, "The Concept of Record in Interactive, Experiential and Dynamic Environments: The View of InterPARES," *Archival Science*, vol. 6, no. 1 (2006), pp. 2-3, 15, 26-28, 32-33.

42 Kenneth Hawkins, "'Set-Aside' in the Semantic Web: Findings and Implications of Other Government Case Studies," presentation, *Seminario Internazionale "I risultati di InterPARES2,"* 13 December 2006, Milan, Italy.

computers, and have the resources for all of this. Such an assumption will render the archival profession incidental or irrelevant in the digital age. On the brighter side, the many activities, developments, and implementations in semantic technologies that speak to the requirements for the preservation and authentication of digital records offer a way for archival science to meet its own challenges and contribute to a Semantic Web that rises above data and information.

One business domain facing these challenges was the subject of the only engineering experiment conducted by InterPARES 2 (IP2) researchers, and the only case study to address the relevance of semantic technologies to the project's aims: to examine the reliable creation and authentic preservation of records in dynamic, interactive, and experiential systems across the arts, sciences, and electronic government.[43] Case study 19 (CS19) "Preservation and Authentication of Electronic Engineering and Manufacturing Records," reported on the effort of three IP2 research partners to develop a new logical preservation format for complex digital objects used in computer-aided design, engineering, and manufacturing (CAD, CAE, and CAM, respectively). Specifically, the experiment used OWL, the W3C specification that extends XML to allow the execution of semantics within metadata schemas, to persist the geometry, topology, and functional characteristics of CAD model objects. The semantic format enabled automated querying of the digital entity's meaning, expressed in its metadata, in order to assess its authenticity.[44] The outcome provides a glimpse of a method for *authentication* independent of proprietary technologies and positioned precisely at the intersection of archival science, diplomatics, and the Semantic Web.

The concept of a logical preservation format highlights the new conceptions of "record" emerging from IP2. One of IP2's findings is that digital records in experiential, interactive, and dynamic systems may not exhibit all the features of a traditional record, and often depend on communications with logically and physically separate entities, services, and instructions to form a whole.[45] A prospective or enabling record stands ready to assist the

43 InterPARES 2 Project, case studies, available at http://www.interpares.org/ip2/ip2_case_studies.cfm (accessed on 26 January 2009).

44 More detail on CS19 than can be given here is available at Electronic Records Archives Program, US National Archives and Records Administration, *Final Report "Preservation and Authentication of Electronic Engineering and Manufacturing Records," InterPARES 2 Case Study 19* (12 September 2005). Compiled by Kenneth Hawkins, PhD, National Archives at College Park, Maryland, available at http://www.interpares.org/display_file.cfm?doc=ip2_cs19_final_report.pdf (accessed on 26 January 2009).

45 Duranti and Thibodeau, *passim*; Luciana Duranti, Jim Suderman, and Malcolm Todd, "Part Seven - Structuring the Relationship Between Records Creators and Preservers: Policy Cross-domain Task Force Report," in *International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records*, eds., Luciana Duranti and Randy Preston (Padova, Italy, 2008), pp. 10-11, available

completion of a business action in tandem with other inputs; the service fee charged to complete an automatic teller machine transaction, for example, is maintained as a fixed entity against which subsequent transactions are run. To the CS19 research partners logical preservation format meant an open, non-proprietary XML-based format that encompassed not only the fixed geometric specifications of the model but also its semantically encoded metadata. The latter - enhanced knowledge about the geometric and topologic features of the model, and the relationships and constraints that characterized them - were first created using proprietary software and ultimately migrated into OWL and there joined to the record to create its archival form. The record has fixed content that memorializes data representing required piece parts and with it actionable metadata that stands ready to enable ongoing automated manufacturing processes. Meaningful preservation requires preservation of both object and ontology.

CS19 was conducted by partners from government and academia: the Electronic Records Archives Program (ERA) of the National Archives and Records Administration (NARA), the San Diego Supercomputer Center (SDSC), and the originating research partner, an agency of the US government (hereafter, CS19 design partner), with responsibilities in the science, engineering, design, and manufacture of complex, high-assurance electro-mechanical assemblies. The CS19 design partner has an ongoing need to access and use its CAD records for business purposes over a long period of time (50+ years) with the assurance that they remain accurate, reliable, and authentic. A broad domain of industrial design and manufacturing firms conducts business using the same software engineering technologies examined in CS19. The global market for CAD/CAM/CAE software applications and software maintenance has averaged ten percent annual growth rates in recent years, reaching $5.45 billion in 2004.[46] Many companies, like airline manufacturers, ship builders, and automakers face the same issues with similar timelines: how to assure the long-term preservation of authentic, reliable records of the complex and proprietary digital entities used to create three-dimensional models of objects, specifying their dimensions, topology, materials, weight, etc., and providing data and instructions processable by the computerized mechanical tools (robots) that manufacture the production piece, part, or assembly.

The activities of the CS19 engineering experiment are given here to help the reader understand the radical implications of its approach: attempting to preserve complex digital records across a domain-specific Semantic

at http://www.interpares.org/display_file.cfm?doc=ip2_book_part_7_policy_task_force.pdf (accessed 1 February 2009).

46 Nancy Wu, "Market Share: Mechanical Applications Software, Worldwide, 2004," 26 July 2005, available at http://www.gartner.com/ (accessed on 26 January 2009).

Web, using open-source XML-based formats for persistence, and a pattern language for authentication instead of authentication based on the examination of media, bit-counts, encrypted "signatures," or the *ad hoc* summing of discrete provenance-based attributes. Each step of the experiment protocol and iteration of the test records format was chosen to either strengthen semantic expressiveness or to capture knowledge representation in a persistent, open source encoding format.[47]

In the business activities of the CS19 design partner's agency, the records set aside include:

1. a Pro-engineer solid model file in CAD native format;
2. a version of the same file in ISO 10303 Standard for the Exchange of Product Model Data (STEP), AP203 format, which precisely describes the boundary representation of a solid model and its formal element, attribute, and behaviour definitions, and Part 21 of STEP, EXPRESS, which gives an object-oriented representation of the object's features as classes;
3. an image of the model in TIFF format.

To meet ongoing business needs, the CS19 agency partner stores these three formats as an aggregate termed the "bill of materials" in a proprietary document management system. The CAD file resides in a proprietary format that requires periodic migrations to stay with its vendor's current versions. The STEP/EXPRESS format is an industry standard that allows files and schemas to be transported across space and, with some loss of features or functionality, to be opened and processed by other CAD programs.[48] None of the practices addresses the long-term archival needs of the records keepers.

In contrast, the scientific activities of the engineering experiment aimed to preserve accurate, reliable, and authentic versions of the record across time, with no loss of features or function, and to minimize reliance on proprietary systems. Test record entities of the STEP/EXPRESS forms were enhanced into, respectively,

4. an encoded format to specify additional geometric relationships and constraints affecting part shape, and action or process semantics to create an "authenticating shape fingerprint." This format was then read into
5. Logistica, a proprietary reasoning engine format to complete a rendition that included the formulation of logical predicates. From this format an extract was derived into
6. the WC3's Web Ontology Language (OWL) format, to complete its trans-

47 The experiment summary is drawn from: *Final Report, InterPARES 2 Case Study 19, op. cit.*

48 Michael J. Pratt, "Introduction to ISO 10303 - The STEP Standard for Product Data Exchange," *Technical Note* (September 2001), National Institute of Standards and Technology, Manufacturing Systems Integration Division, available at http://www.nist.gov/ (accessed on 26 January 2009).

formation into an open source, public domain XML specification for persistent archiving purposes.

To get an idea of the type of solid model and the features specified in the logical preservation format, consider the relatively simple object in Figure 2.

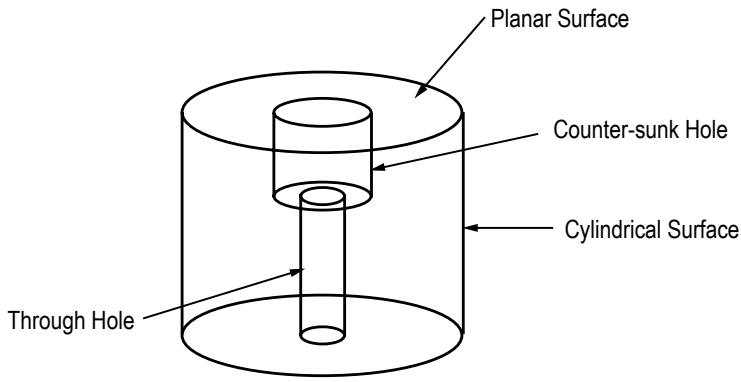

Figure 2: Features of Shape

Each face, plane, radius, hole, and surface has different characteristics, including measurements, tolerances, and purpose. All features, including exact measurements, and the relationships between them, were specified precisely as classes in the artifacts of the CS19 experiment. By reasoning across the definitions of classes and their allowable relationships specified in the OWL metadata to the fixed data representing the solid model, the experiment was able to infer whether the XML representation of the model aligned with the specifications and semantics of what such an object was allowed to be within the closed world of the experiment's pattern space (the navigable space between the solid model object and its corresponding ontology).

Another example of this approach that was not part of the experiment may make it clearer. Reasoning across the attributes and relationships of the features of the object in Figure 3 allows an inference to be made as to the identity and integrity, and therefore, the authenticity of a model representing a real world object.

Figure 3: Solid Object Model

- Shape: spherical
- Diameter: 1.68”
- Surface: dimples
- Gravity: centred

Guided by an OWL ontology representing the characteristics of classes and the relationships constraining them in the domain, a reasoning program would be able to examine the semantic expressions encoded in an object file purporting to be authentic to this domain, and infer if in fact it was what it purported to be: a CAD model of a golf ball. If the values in one field contradicted the ontology (“smooth” instead of “dimples” in the field for Surface) the method would not offer a presumption of authenticity. The simple correlation of discrete attributes using logical interrogation would allow a computer agent to infer authenticity and, based on business rules, make the record available prospectively in support of subsequent actions or procedures. In a CAD/CAM environment that automatically produced a variety of products, this could mean changing a manufacturing procedure from one product to another.

Precise specifications of part shapes, class cardinalities, and some but not all of the relationships deemed necessary for a presumption of authenticity, were successfully transformed and reasoned over using CS19’s experiment protocol and tools. The limits of OWL’s expressivity for capturing in logical statements knowledge representation of the business object, however, meant that it could not capture the “action semantics” of the part within its manufacturing context. Logistica, the proprietary reasoning engine used in the experiment, was better able to express these elements but they remained trapped in it along with some of the run-time processes needed to actually interrogate them. As the CS19 design partner noted, “we cannot put all of the knowledge about the part in archival form.” And to produce the part would require moving the archival form back into a proprietary CAM environment. While the power of open semantic encoding schemes is constantly being improved by the architects of the Semantic Web and others, the issue of proprietary software highlights a pervasive challenge: “preservation is in direct conflict

with private industry."[49]

The digital objects were not considered authentic by the exacting standards of the CS19 experiment partners, because the entire range of semantic metadata representing essential patterns within them and their business context did not remain available when the test concluded.[50] Like all engineering experiments, however, failure to succeed completely does not preclude future successes nor, in CS19, diminish the insights won in the attempt.

A thoughtful review of CS19 and its technical and historical context within OO software engineering, highlight the opportunities and risks faced by archival science and diplomatics today, and into the future. To begin with, the findings of CS19 directly address IP2's research agenda to investigate how other domains, in this case science and electronic government, understand questions of authenticity, reliability, and accuracy when applied to new record types and records aggregations in systems with interactive, dynamic, and experiential characteristics. IP2 also called for a translation of domain findings to archival science and diplomatics where each could inform and benefit the other. CS19 did this by showing how concepts, methods, and tools developed by advocates of the pattern language approach to software architecture and engineering have opened the possibility of a new methodology for the discovery, authentication, and preservation of electronic records before, during, and after their transition from active use to archival preservation. That is, it attempted an archival version of the Alexandrian/OO response to complexity and volume: "thin-slicing" the most salient characteristics of the patterns at work in a particular (digital) space, and making them explicit and available to automatic querying by semantic tools for authentication, work, and preservation.

The examination of authenticity, reliability, and accuracy of digital records "as they are understood in the various disciplinary areas involved in the research" was the second of three domains across the three focus areas of IP2 and one that proceeded from the findings of IP1 on these topics.[51] (The emphasis on how "various disciplinary areas" understand these concepts is a critical one to understanding CS19's implications and we will return to it

49 CS19 design partner to author, 16 February 2005.

50 *Final Report, InterPARES 2 Case Study 19,* p. 19.

51 "Overview of InterPARES 2 Intellectual Framework," available at http://www.interpares.org/ip2/ip2_intellectual_organization.cfm (accessed on 26 January 2009); John Roeder, Philip Eppard, William Underwood, and Tracey P. Lauriault, "Part Three - Authenticity, Reliability and Accuracy of Digital Records in the Artistic, Scientific and Governmental Sectors: Domain 2 Task Force Report," pp. 38-39 in *International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records*, eds., Luciana Duranti and Randy Preston (Padova, Italy, 2008), available at http://www.interpares.org/display_file.cfm?doc=ip2_book_part_3_domain2_task_force.pdf (accessed on 26 January 2009).

momentarily.) InterPARES 1 drew a distinction between authenticity and authentication, as is necessary, defining authenticity as a quality had by a record "that is what it purports to be and is free from tampering or corruption." It then stated that "common usage" defined authentication as a time-specific attestation of a record's authenticity by a trusted juridical person with authority to so state in the form of an annotation (or attachment) made to the record. The prevailing technological means of making such an annotation to an electronic record, the so-called "digital signature," was rightfully deemed as insufficient based on its narrow approach (and its reliance on proprietary software).[52] "Authentication does not establish authenticity over time."[53] IP2 afforded the opportunity to demonstrate that authentication could employ technologies consistent with archival and diplomatic concepts to assert the authenticity of an electronic record without altering it or simply counting its bits.

For the reasons given by IP1 and detailed throughout this article, reliance on single-technology authentication "solutions" based on media checks, bit-counts, encrypted seals, and worse still those that are implemented by proprietary software, is not sufficient. After providing a detailed briefing on non-monotonic logic, theorem provers, and the need to preserve, using open standard encoding specifications, "the enduring reasoning process" and "the enduring reasoning form" present at records creation in order to assure viable digital authentication, the CS19 design partner quipped:

> Authentication based on media (are all the bits in the same place and have any of them changed?) is quite easy by comparison. The heart of the issue is that *there is considerable knowledge that does not exist in any digital form that is crucial to the preservation of digital objects*.[54]

The remark, made in the context of a complex engineering experiment in a domain external to records, led and conducted (mostly) by experts from a field other than records, makes it clear that IP1's concepts of archival science and diplomatics – sensibly combined with technology – helped advance a viable new approach to digital authentication. Its second point goes to probably the most significant insight IP2's CS19 gives about the requirements of a method for authenticating and preserving electronic records in dynamic, interactive, and experiential computing environments. The basis for ascertaining quality in software engineering and authenticity in an electronic

52 *ATF Report,* p. 2.

53 Luciana Duranti and J.-F. Blanchette, "The Authenticity of Electronic Records: The InterPARES Approach," in *Proceedings, IS&T 2004 Archiving Conference*, 20-23 April 2004, available at http://polaris.gseis.ucla.edu/blanchette/papers/ist2.pdf (accessed on 26 January 2009).

54 CS19 design partner to author, 16 February 2005 [emphasis added].

record lies in detecting and making explicit the identity and relationships between classes, and the characteristic logical patterns that hold them together. The patterns affecting the archival undertaking are precisely those in which the attributes of provenance and creation participate. In terms of establishing sufficiency for either a presumption or verification of authenticity, the findings of CS19, backed by the Semantic Web concepts and tools and, to an extent, the findings of IP2 in regard to meaningful distinctions between memorial and enabling records,[55] advanced IP1's statements about implementing its "Requirements for Assessing and Maintaining the Authenticity of Electronic Records" to their next logical step: enlisting automated tool support for the processes archivists can no longer implement effectively through human actions.[56]

IP1's Authenticity Task Force noted that its baseline requirements apply to the activities of the preserver, acting as a trusted custodian, and referred to different approaches of implementing the requirements. The Task Force also expressed concern, however, about the preserver's workload in assessing the volume of data necessary to reach a presumption or validation of an electronic record's authenticity. CS19 therefore tested the proposition (already operational in the Semantic Web) that to depend on human agents to enumerate discrete identity and integrity metadata is inadequate to the demands for discovery, preservation, and authentication facing archives now and into the foreseeable future. The listing and keeping of these attributes from the creator's custody onward is critical but the work carried out for preservation must do more with this kind of information. The conception of intrinsic documentary form needs to go much further into recognizing the characteristic patterns (classes, relationships, constraints) that cohere among and between otherwise discrete identity attributes. The correlation of attributes signifying business context is essential to finding these patterns.

If the character of a place, as Alexander insists, is conferred on it by the episodes that happen there,[57] the character of a record is given to it by the episodes that happen around it. The patterns that emerge as the true documentary forms that we already know are the repeating traces of the repeating actions that give rise to a fonds of records. The patterns themselves have embedded rules that describe the way they can be created and of what they may consist. The patterns that repeat are always anchored to a form in space and give the building (or documentary form) its essential character. Authenticity is supported upon the logical correlation of one characteristic element or attribute of an object with one (or more) other characteristics

55 Duranti and Thibodeau, *passim*.

56 *ATF Report,* pp. 20-23, and Appendix 2, "Requirements for Assessing and Maintaining the Authenticity of Electronic Records," *passim*.

57 Alexander, *The Timeless Way of Building*, pp. 62, 183-85.

elements or attributes, in accordance with known criteria (such as ontologies, class relationships, business rules, etc.). Instead of relying upon the *ad hoc* examination by human agents of discrete identity attributes, the CS19 engineering experiment joined these concepts with semantic technologies currently in wide use across the World Wide Web (in government, science, and commerce) to support the automated discovery, authentication, and use for work of the patterns and correlations of patterns present in the expanding volumes of digital records. When organizations spend millions of dollars designing CAD models jointly with procedures for manufacturing purposes, and generate millions of such records needing preservation for decades, it is imperative that the archival profession move beyond work processes requiring one-on-one interactions by individual archivists for discovery, authentication, use, and preservation. To fail here would mean to remain working at the lower levels of the Semantic Web, examining *ad hoc* merely for the absence, presence, or sum of discrete identity and integrity metadata of small data sets, one transfer at a time.

At its conclusion, IP1 determined that its efforts to create record templates and typologies of documentary forms that cut across domains (the approach of general diplomatics) should be supplemented by methods focused instead on the records systems and aggregations that were specific to organizations, domains and juridical systems (the approach of special diplomatics and archival science), and build from there to the general. "Increasing the utility of diplomatics as an aid to understanding diverse electronic systems will require the development of a more nuanced interpretation of the characteristics of electronic records and the manner in which they manifest themselves in a variety of electronic environments."[58] In this way IP2 followed, albeit subtly, the same Alexandrian pattern as had the OO, system modelling, and Semantic Web communities. CS19 explicitly employed the tools and concepts of the pattern language approach used throughout OO software engineering, and related architectures and semantic encoding formats, where the repeating patterns of events, activities, and requirements peculiar to the business domain at hand are essential to devising technical solutions that address them meaningfully. IP2's Domain 2 authors concluded that CS19 was among the studies that "confirmed the need, suggested by the conceptual analysis, for expansions to the traditional conceptions of authenticity, reliability, and accuracy."[59]

This approach holds promise for discerning and making use of the pattern languages in documentary forms to build bridges from special diplomatics to logical preservation formats. As one author noted, "What's needed is the

58 *ATF Report,* p. 24. See also pp. 14-16.

59 Roeder, Eppard, Underwood, and Lauriault, p. 39.

recognition that the semantics of a schema (or, more precisely, the semantics of data governed by a schema) must be explicitly bound to a known community that it serves, and that bridges between the communities will be an inevitable part of any comprehensive solution …"[60] Mapping the "navigable space" between the patterns reoccurring in the domain and the OO system elements to harness them for work, establishes the reference points against which future semantic interrogation of electronic records can proceed. The method enables finding the correlations between, rather than summing the number of, provenancial and procedural attributes. It does not diminish but finds kindred spirit with diplomatics, as Luciana Duranti established in her essays on the relevance of contemporary archival diplomatics to the preservation of authentic electronic records.

> Briefly, where records creation is consciously controlled, diplomatics guides the recognition of patterns and facilitates identification, while, where records creation is uncontrolled, diplomatics guides the establishment of patterns, the formation of a system in which categories of records forms are devised, which is able to convey content and reveal procedure. Once a system is established, then its description in a metadata system will have to reflect it by expressly articulating the relationships among record forms, procedures, actions, persons, functions, and administrative structures.[61]

CS19, along with at least two other IP2 case studies, demonstrated that highly specialized metadata related to a specific domain, discipline, or business activity need to be understood and translated to ensure the preservation of authentic, reliable, and trustworthy digital records.[62] These metadata make explicit, at the record level, the patterns that connect the classes and objects of systems that are used to meet domain-specific business functions to their creating context. Although sometimes overlooked in the desire to create tools within archival science based on the notion of a canonical record, much of the data and information necessary to preserve meaningful records and enable viable authentication using semantic approaches is embedded within domain-specific "resource description" schemas. These help situate records within their creating context, and reveal the provenancial attributes most important to the business owner and how they relate together in managing real records.

60 William C. Burkett, "The Myths of 'Standard' Data Semantics. Faulty Assumptions Must Be Rooted Out," *XML Journal*, vol. 3, no. 11 (November 2002), quoted in "XML and the Semantic Web," at http://xml.coverpages.org/ (accessed on 26 January 2009).

61 Luciana Duranti, "The Uses of Diplomatics," in *Diplomatics: New Uses for an Old Science*, p. 175; see also, pp. 30-32*ff.*

62 CS06, Cybercartographic Atlas of Antarctica, and CS18, Computerization of Alsace-Moselle's Land Registry. The framework and methodological approach of IP2 itself emphasized interdisciplinarity and transferability. See "Overview of InterPARES 2 Intellectual Framework," pp. 3-4. See also Roeder, Eppard, Underwood, and Lauriault, p. 39.

The tension between the agreed-upon basis for fixing in instruments and then making explicit the patterns in a domain, has allowed considerable progress in the engineering disciplines, both of built spaces and those addressed by object-oriented software design and development. Applied to archival science and diplomatics, the establishment of a "navigable space" promises to work as well for electronic records in traditional documentary forms as those more complex forms addressed here. The opportunity of the business expert in archives today is to join the discussions currently underway within the technical communities implementing Semantic Web technologies, whether in the arts, sciences, government, or combinations of each. That is, the archival profession should be doing more to help identify and model the patterns manifest in documentary forms specific to the discrete business domains in which they currently work. For example, even preliminary investigation shows that many business domains are either actively developing or need to develop useable RDF and OWL ontologies, the input to which archival thought, support, and participation would be welcome.[63]

The identification of the most telling relationships between classes and their representation in the attributes joined to business objects of this kind are already proceeding, thanks to the willingness of the architects and builders of the Semantic Web to borrow from the archival profession. In 2006, the Object Management Group approved the development of an industry-standard specification for records management services by its membership, which includes private sector companies, academia, and government. The Joint Records Management Specification (JRMS) activity is based on the functional requirements, use case, and UML models for the Federal Enterprise Architecture developed by the eighteen largest-funded US federal agencies in conjunction with the ERA Program at NARA. Records management services will capture the context of creation at the point of creation and carry it forward, updating as necessary, adding management attributes, and providing management services through the entire record life cycle.[64] The JRMS activity is the first time a set of requirements derived from archival science and diplomatics - generated and endorsed from experts in civilian and defense agencies - has moved into an industry standards process. The OMG requires vendors developing the specification to implement it in their products within

63 See "Semantic Web Case Studies and Use Cases," W3C Semantic Web, available at http://www.w3.org/2001/sw/sweo/public/UseCases/ and "Catalog Of OMG Domain Specifications," Object Management Group, available at http://www.omg.org/technology/documents/domain_spec_catalog.htm (accessed on 26 January 2009).

64 *Functional Requirements, Attributes, and Unified Modeling Language Class Diagrams for Records Management Services* (7 September 2006), available at http://archives.gov/era/rms (accessed on 26 January 2009). The JRMS includes functions for record capture, provenance, category (archival bond), authenticity, case file, disposition, and reference. For the integration of RMS into the FEA, see http://core.gov/ (accessed on 26 January 2009).

two years of approval, which is currently on track for mid-2009. In addition, the effort is seeking participation from experts, implementers, and end-users as this work goes forward.[65] The JRMS is not the only example of archival science furnishing concepts and requirements for integration into Semantic Web architectures. Computer scientists and engineers working on extensions to the functionality of the Resource Description Framework (the W3C specification that makes data and information discoverable on the Web within the context of its relationships to associated objects), are building the concept of provenance into the specifications for RDF to enhance its expressivity for trusting Web resources.[66] The archival profession should embrace these opportunities and identify additional ones to carry forward a true collaboration with the business and technical communities.

Conclusion

Whether the path is taken by joining the development of technical standards and specifications, working to document and operationalize patterns within business domains, or providing input into RDF and OWL specifications for use by a single organization or group, the central responsibility of the business expert in archives is this: identify the characteristic patterns that cohere among, between, and within the classes of the business space of concern, or at least consider the potential of a technical approach that would make what was once latent in the archival method explicit and render it available to automated tools. Then, collaborate on the specifications or requirements based on these patterns together with the system and data communities now developing systems whose records are threatening to overwhelm us. When developed gracefully, with quality, the systems will emanate from these same patterns, will fix those patterns permanently in the documentary entities of systems development and operation, and will rely upon them long before any code is constructed and long after they enable the preservation of authentic electronic records.

65 Hawkins, *op. cit.* For the status of the JRMS specification at the Object Management Group, see http://gov.omg.org (accessed on 26 January 2009).

66 Carroll et al., *op. cit.*; Edd Dumbill, "XML Watch: Tracking Provenance of RDF Data" (21 July 2003), available at http://www-106.ibm.com/developerworks/ (accessed on 26 January 2009); Jun Zhao, Chris Wroe, C. Goble, R. Stevens, D. Quan, and M. Greenwood, "Using Semantic Web Technologies for Representing e-Science Provenance," Proceedings of the 3rd International Semantic Web Conference, *Lecture Notes in Computer Science*, vol. 3298 (2004).