4. **Puddling – tools and approaches to selective assembly of digital information content in support of curated online collections and/or specialized information services**

    4.1. How can Aquifer leverage existing taxonomies to improve search capability and/or to assist in the automated selection (curation) of digital collections that are based on large-scale aggregations of materials assembled via OAI harvesting, web crawling, or other means? What taxonomies exist that could be useful? What are the technical obstacles that need to be overcome so they may be utilized effectively, for example, as web services?

The charge obliged us to *develop a detailed specification or problem statement for work in this area along with recommendations for funding required*. More realistically, let's peg this as an effort to scope the problem space – after which Aquifer can pursue detailed specs at a level that can drive useful work plans and funding estimates.

Business community

Staying within the water analogy, this summary surveys the results of dropping a *taxonomies-pebble* into the WWW pond, following the ripples for a few of days, and thinking about what came into view. Not surprisingly, semantic-web references showed up early and often. Rather than focus too narrowly on that community's activities at the outset, it seemed worthwhile looking in venues that might provide indications of where the information industry and information services embedded in corporations are headed.

Terminology in analytical/predictive services aimed at the for-profit community does in fact include references to the semantic web:

> A decentralized approach to X-internet interoperability will win vendors and users will adopt a new, evolutionary approach to interoperability. Based on Semantic Web research, decentralized data dictionaries will let companies publish how their products communicate. Tens of thousands of these dictionaries will emerge as technology suppliers and users base new efforts on the most successful ones, regardless of who created them. This will cause consolidation toward de facto standard dictionaries by 2010.[1] *[any questions !?]*

Looking a little deeper, we can find architectural maps that have lots in common with the environments that we're building to drive programs and services for our institutions. A pair of diagrams from a business analysis document from early this year illustrates the similarities clearly (appended as last page). And a more recent piece includes some thought provoking ideas about the characteristics of future knowledge-management services for businesses in the form of three types of service activity:

[1] David Truog, How The X Internet Will Communicate [The Forrester Report], December 2001

- Abstract … construct and expose new composite views of information [based on] instrumentation to extract or derive metadata across a continuum of structure [ranging] from no structure, [thru] XML descriptions [and] text concepts [thru] classifications to relational schemas and views …
- Adapt … recognize, accommodate, and retain a user's context and information access rights, as well as changes in the data itself; -- … detect relevant changes in the associated content and synchronize indexes and caches; -- … propose updates to the virtual metadata library based on newly encountered data or incoming requests …
- Respond … hide the underlying data complexities from the user; -- … federate the request, handle any authentication requirements of data or content sources, and identify the user's context; -- … translate the request into multiple requests than can be filled directly (by search engines, for example) or passed to integration middleware software that can transform it …

This study and report by Forrester was based in interviews with the following companies:

> Access Systems Consulting, Bearing Point, ClearForest, Endeca Technologies, Hummingbird, IBM, Identitech, Intelliseek, Mobius Management Systems, Oracle, Pegasystems, Pervasive Software, Plumtree Software, SAS Institute, Teradata, Venetica, Verity, Vignette

As such, it provides validation that a fair number of for-profit endeavors are looking at important aspects of semantic-web like services.[2]

Furthermore, spending a bit of time in the business analysis and reporting arena provides some reassurance that a goodly number of companies are working actively toward these types of services. Among others who appeared in various listings and summaries are:

> Attensity, Metacarta, Kofax (Mohomine), NovoDynamics, SDR, Stratify, Tacit Knowledge Systems, Traction, and Zaplet, as well as SAIC and SRA.

XML activity

A second venue – XML activity in general -- presented itself as the ripples continued to spread. An interesting view of what's going on currently was available in the papers from *XML Europe 2004*.[3] Here there are half a dozen substantive papers related to the semantic web. And we also find another half dozen focused on issues associated with topic maps.

Topical mapping provides a framework that can support definitions of relationships among metadata elements. For example one could define the utility and scope of

---

[2] DLaurie M. Orlov and Laura Ramos. Organic Information Abstraction [Forrester Big Idea], May, 2004

[3] http://www.idealliance.org/papers/dx_xmle04/
also take a look at:
http://www.idealliance.org/resources/proceedings.asp
http://www.idealliance.org/resources/peers.asp

resources in relation to their reliability, or their brevity, or their inclusion of access to the proofs on which was publication was based.[4]

And, within one of the XML Europe papers,[5] we find a very interesting study:[6]

> EP2010: The Future of Electronic Publishing Towards 2010. A strategic study on the future of research into publishing, content and knowledge technologies. Authors: Wernher Behrendt, Guntram Geser, and Andrea Mulrenin, all of Salzburg Research. September 2003: European Commission Directorate-General for the Information Society.

Among several useful documents here is one focused on *Smart Content* in which the authors elaborate their objectives as follows:

- Ideally, the properties of Smart Content are such that the consumer begins to associate tangible qualities and benefits with "smart" as opposed to "dumb" content.
- The vision is to focus the research activities in a way that fosters standardization to gain direction and technological innovation to gain economic momentum.
- The major function of the "Smart Content" concept is to stimulate and support the discussion of novel digital content technologies or solutions and - if applicable - of associated novel value chains

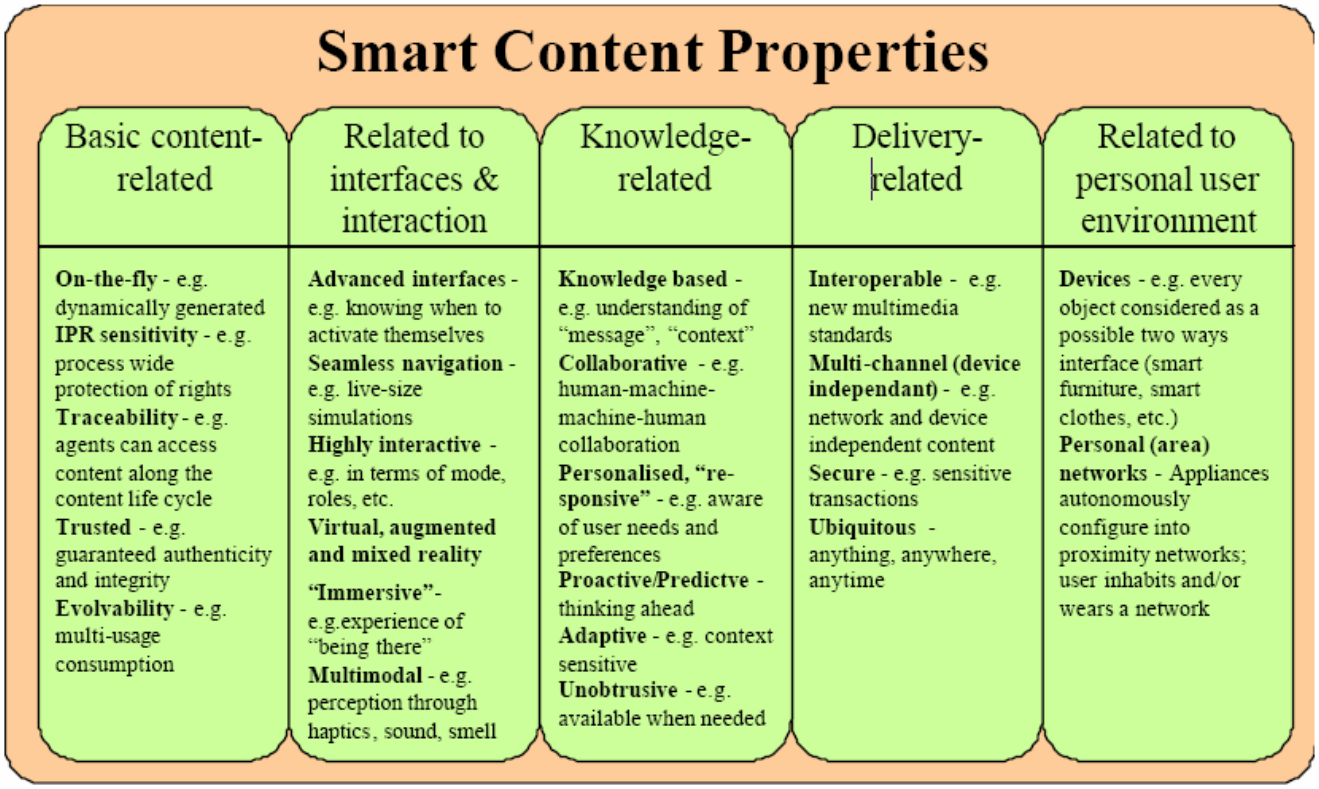
They go on to provide a chart-like summary of the attributes such content would exhibit:

---

[4] http://www.y12.doe.gov/sgml/sc34/document/0446.htm
http://xml.coverpages.org/ni2004-04-09-a.html
http://xml.coverpages.org/topicMaps.html (1 of 85)21/10/2004 2:26:39 AM

[5] http://www.idealliance.org/papers/dx_xmle04/papers/04-01-02/04-01-02.pdf

[6] http://ep2010.salzburgresearch.at/



Semantic web

With the above bit of real-world validation (and a small side trip toward 2010) in hand, we come then to the present and near-term state of affairs associated with the Semantic Web framework.[7] Among the areas of special activity that are germane to Aquifer's objectives are the efforts found under *Advanced Development.* Also worth a look is what's laid out under *Best Practices and Deployment.* Note too that *RDF* remains very active.

Of special importance to our efforts is the recently completed work for the Web Ontology Working Group and *OWL Web Ontology Language:*[8]

> 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. OWL can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms. This representation of terms and their interrelationships is called an ontology. OWL has more facilities for expressing meaning

[7] http://www.w3.org/2001/sw/
[8] http://www.w3.org/TR/owl-features/

> and semantics than XML, RDF, and RDF-S, and thus OWL goes beyond these languages in its ability to represent machine interpretable content on the Web.

As with topic maps, OWL brings a processing framework into play as the means to manage and deliver the capabilities and benefits that can be derived from taxonomies (as well as other guides to content).

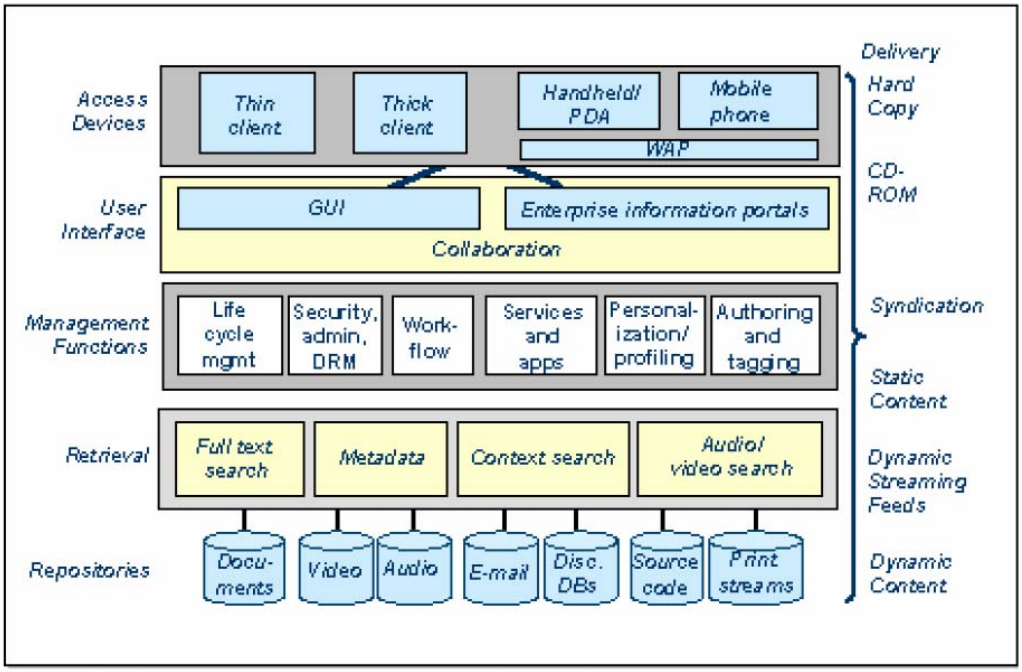
Summary

This effort is obviously incomplete … many important aspects of this arena remain untouched:

- identifying and scoping the extent and utility of extant taxonomic efforts
- creating and testing models for tools, effort, and costs that would go into building and managing topical analysis for varied categories of resources
- investigating other technologies for clustering results, for presenting results in more intuitive/effective ways, for delivering pervasive access across institutional, program, and service boundaries

The effort here was consciously focused on framing the problem space in such a way that projects in this arena could be identified, spec'd and launched within the broader spectrum of the semantic web framework. Important as the metadata is -- whether it be OAI-DC, or machine-assisted taxonomies, or new constructs that provide richer, deeper access to content -- the end product of our efforts needs to be a user experience that satisfies a real need. It seems apparent that the semantic web framework's objectives could help us define objectives for our efforts that will put the needs of faculty and students up front. We can always find ways to "do repositories better" or "do content sharing better" or "do metadata better". First though, we might want to find ourselves a very few, very visible projects that will put digital content to work in ways that improve the academic productivity of our faculty and students. In other words, it's probably past time to put a little pizzazz in the water.

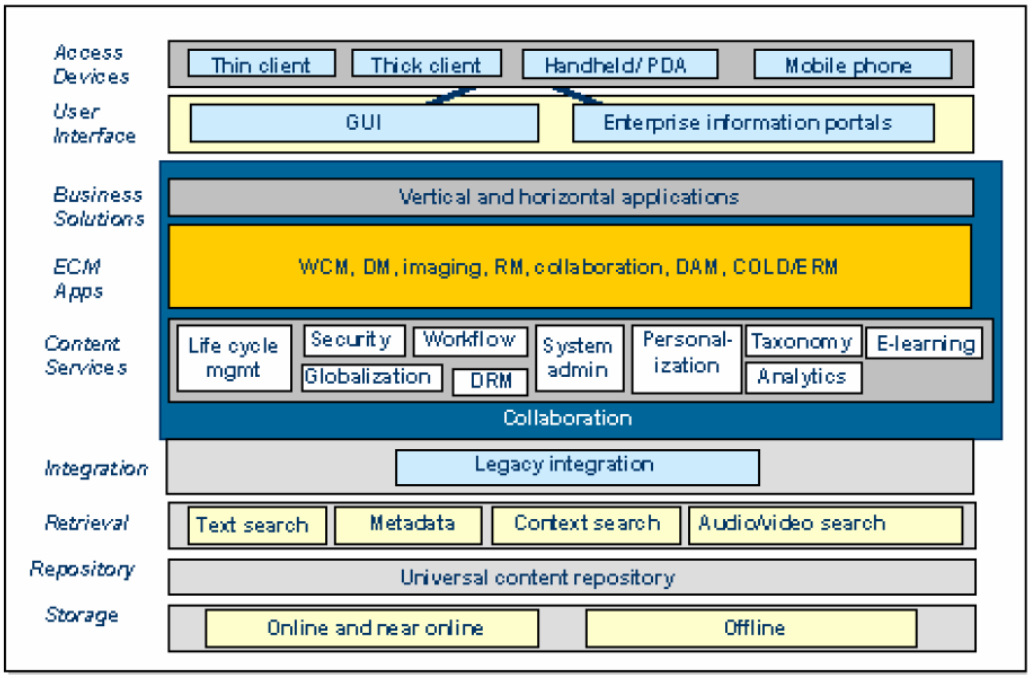
Figure 1: Original Enterprise Content Management Architecture (2000)



Source: Giga Research, a wholly owned subsidiary of Forrester Research, Inc.

9

Figure 6: ECM Architecture for the Repository Infrastructure Phase



Source: Giga Research, a wholly owned subsidiary of Forrester Research, Inc.

10

[9] Connie Moore and Robert Markham. Will the Real Enterprise Content Management Please Stand up? [Giga Position], January 2004

[10] *Ibid.*