E. Duval, W. Hodgins. *Metadata Matters*, Submitted, June 2004. See also: http://rubens.cs.kuleuven.ac.be:8989/mt/blogs/ErikLog/archives/000566.html

Metadata Matters

Erik Duval
Dept. Computerwetenschappen, K.U.Leuven, B-3001 Leuven, Belgium
Tel. +32-16-32.70.66
Fax. +32-16-32.79.96
EMail:erik.duval@cs.kuleuven.ac.be

Wayne Hodgins
Strategic Futurist, Autodesk Inc.
Tel. +1-415-507-5759
Fax. +1-707-773-1285
EMail:wayne.hodgins@autodesk.com

Abstract: In this paper, we argue that, in order to facilitate the ubiquitous uptake of metadata, and in order to realize their potential for advanced flexible end user functionalities, the metadata should become more invisible for the end user. We also argue that we are technically capable of realizing this goal, and illustrate the issues involved with some practical examples.

Keywords: learning objects, learning object metadata

1 Introduction

Having spent so much of our time and effort over the last few years on the development of concepts, standards, tools and infrastructures for metadata, the authors are more than pleased with the steadily increasing attention and focus on this topic, both in the R&D world as well as in the commercial marketplace and not-for-profit sphere.

However, we believe that we are still very early on an exponential curve and that many of the current developments and efforts are somewhat misguided: in our view, they often place too much of an emphasis on the elusive quest for perfection and thus illustrate that "the perfect is the enemy of the good"[1]. Perhaps even more disconcerting is our concern that many of these efforts are perfecting the irrelevant, as they are they focus on the literal use of metadata, thus seeking to continue historical and current practices, rather than trying to design, experiment with and implement more innovative and effective ones. Moreover, many current developments do so at the expense of end users, who are supposed to spend considerable time and effort on the definition of detailed metadata, using obscure terminologies and unreadable, machine oriented syntaxes.

As the authors have been, and continue to be, deeply involved in standardization activities around Learning Object Metadata[2], we want to point out that standards are meant to enable developers to realize interoperable technical components. Standards are not meant to be visible to end users! As an example, too many tools and implementations of LOM and SCORM use the exact same terminology as in the LOM standard document. Terms like "catalogue entry", "contribute" and "semantic density" are fine and appropriate for the standard document itself. However, these terms are not likely to be terms that are familiar and understood by many audiences and communities who are creating and using such metadata. It is understandable that early implementations of new standards and

[1] French proverb.

[2] IEEE Learning Technology Standards Committee (LTSC), Learning Objects Metadata (LOM) http://ltsc.ieee.org/wg12

specifications focus on the implementation of the functionality required. However, it is obvious that evaluations of the actual experience of end users with these tools will show the failure of this approach.

Indeed, just like web browsers do not disclose the hairy details of HTML or HTTP, usable tools should not expose detailed Learning Object (or other) Metadata. We should hide those details and develop tools that do not unnecessarily burden or complicate the life of the end user.

These observations are meant to assist us in moving on to the next stage, rather than to be merely critical or negative. Just as with the early days of the Web, when the only (or at least most widely used) tools available were text editors, and authors provided the actual HTML tags to create HTLM documents, we can see how the evolution of HTML tools has proceeded to the point of making this largely transparent to end users, content creators and readers alike. In the usual clarity that hindsight provides, even the very title of "HTML editor" exposes the risk of "perfecting the irrelevant". Nor is it any coincidence that the tipping point of the exponential rise in the use and benefits of HTML matched this evolutionary path of tools which masked the underlying complexity of the HTML standard and let creators and consumers keep their focus on the content. XML is now following a similar route and the intent of the authors here is to promote the most rapid migration to this next phase of metadata implementation. The goal is to make the use of metadata similarly easy and transparent. This is the way to achieve widespread adoption of metadata and reap the inherent benefits

In this paper then, we try to address some misconceptions and myths with regards to metadata, in the hope that this can help to focus R&D on approaches that will result in powerful flexible enablers for end users. In this sense, the paper can be regarded as a follow-up to [3].

2 Basic Message

The basic message of this paper is that, in order to be successful, metadata should become invisible for end users. This is just a specific instance of the general observation that technology can become successful only when end users become unaware of its presence [Norman, 1998]. Indeed, few people are aware of the "user interface" to tools like cars, telephones, etc. Such tools do not confront us with the complexities of the underlying technologies and infrastructures.

Because metadata is still too "visible", there really isn't very much penetration of metadata in the light of the enormous amounts of digital content out there, despite the uptake and adoption of metadata by many groups. Moreover, many of the metadata are in the form of "file properties" or locked in systems such as LMS, LCMS, ERP, CMS, etc. Most of the metadata that reside in these systems remain unexploited.

The authors do not just want complain about this situation or bemoan it. Rather, we want to emphasize the need for continued awareness raising of the power of metadata, NOT by putting a focus on metadata itself, but rather by emphasizing what is possible when metadata is present. This includes demand creation for metadata, clear understanding of the business case for metadata, lessons learned, and recommended practice.

Moreover, even though metadata should remain invisible for end users, we do believe there is a serious need for large communities of practice; professions, etc. to take on the responsibility of providing the critical elements of relevance and applicability to their disciplines and constituents. This includes the hard but necessary work of defining vocabularies, taxonomies, ontologies, etc. Referring again to the example of the telephone, it is largely invisible and transparent to use, only because groups of people spend long hard hours working on such things as not only a common dial tone, but also on working out the agreements between the telephone companies on how to use each others infrastructure, how to seamlessly pass calls back and forth across their systems, interoperable and common phone number systems globally (well almost!) and sorting out the immensely complex cross billing issues to create simple single billing for customers.

In this context, it is crucial to understand that early adopters (and as a reader of this paper, chances are extremely high that *you* are one of them) are *not* representative of the majority of users! Indeed, in the field of Human-Computer

Interaction, the term “crossing the chasm” is used to refer to the fact that the end user experience of tools needs to be transformed in dramatic ways in order for such tools to make it into “mainstream” use!

More specifically in the context of metadata then, this paper argues that:

1. Content in general, and learning objects in particular can be described in considerable detail with acceptable quality through automated means, basically by exploiting the context of use and readily available information about the users involved. Section 3 elaborates on this and the myth that all metadata must be manually created.
2. Complementary to the above observation, we do believe that there is ample room for manually created metadata. Section 4 focuses on the importance of context and community in order to make manual creation of metadata a “labor of love” rather than the burden it currently is.
3. Most content objects will have multiple sets of metadata associated with them, especially as metadata, particularly some of the most powerful, can be quite subjective. Section 5 will argue that “this is a feature, not a bug”. A nice illustration of this concept and its potential to enable truly personalized learning is the way that the TiVO personal video recorder functions.

3 Metadata can be generated automatically

3.1 Introduction

There is a widely spread belief and assumption that metadata can only be provided by humans, preferably with a professional background in indexation, as in the case of library cataloguers. While we would want to have as much of this type of metadata as possible, it is readily apparent that such a manual approach can never scale and that this approach would imply a requirement to deposit all documents intended for “publication” on the Web in a queue for a professional cataloguer. Therefore, in addition to encouraging the continued creation of such “professional metadata” we propose and illustrate the ability to augment this with the infinitely scalable model of “mass contribution” of metadata from “the rest of us”.

We will illustrate in this section that a large number of sources for such mass contribution of quality metadata are readily available and that these metadata can be combined with automatic generation of metadata in a pragmatic way. Nor does this require any breakthroughs in artificial intelligence or other advanced techniques for content processing! (though these are welcomed and encouraged)

3.2 Document itself

A first and obvious source of metadata that can be automatically generated is the content we want to describe itself.

Of course, we can and should apply Artificial Intelligence and related techniques in as far as they have become practically usable. As an example, document clustering techniques now offer viable approaches to group documents together. An interesting application of this kind of technologies is employed in tools such as Grokker [www.groxis.com], and Kartoo [www.kartoo.com]. Such tools are all the more interesting because they provide a visual approach to manipulate the resulting document sets, rather than an electronic form model.

Besides such more sophisticated techniques, we can often use quite simple and pragmatic approaches to enrich metadata:

- For instance, it is a relatively simple process to extract from an HTML document the title, the language used, references to other documents, the name of the author (often included in the metadata that are inserted by the authoring tool), etc.
- “HTML scraping” techniques are used by RSS and Blog aggregation tools, newsreaders, etc. to harvest metadata from existing content.
- It is not so difficult to determine the language for textual documents. Typically, looking up 20 words or so in a number of dictionaries suffices to make a very reasonable guess.
- ActiveX components enable the extraction of the title, author and other metadata from MS-Office documents.
- Etc.

Search engines like Google illustrate that existing harvesting techniques can be quite powerful. (Indeed, we believe that the success of Google is to a large extent due to the smart ways in which it generates metadata automatically, a posteriori, and to the equally smart ways that it succeeds in hiding the complexities of for instance the PageRank algorithm it based much of its result ranking on!) More research on mapping the results of such automated techniques to the different metadata elements in structures such as LOM or DC is needed.

3.3 Context of use

Besides the content itself, the authoring *context* can also be exploited to harvest metadata in a way that can be largely transparent to the end user. If a metadata authoring tool is launched within the context of a course for instance, then it seems reasonable to suggest metadata of the course as starting values for the metadata of the LO.

As an example, in Leuven, we have integrated the ARIADNE learning object repository, called the "Knowledge Pool System", with the Blackboard Learning Management System, and we have been able to capture a detailed set of Learning Object Metadata, without requiring the end user to provide these metadata manually! Rather, we mine the data already present in the administrative system of the university, we exploit the context of operation and information about the user, etc. to deduce all the relevant data automatically behind the scenes.

This kind of work can readily be generalized. Most Learning Management Systems, embed learning objects in the context of a course. Typically, there is a lot of information available about such a course, including

- the audience, for instance "2nd year engineering students";
- the language of instruction, for instance "German";
- the subject of the course, for instance "introduction to relational databases";
- the difficulty level, for instance "advanced placement";
- the learning time allotted to a particular learning object, for instance 20 minutes;
- etc.

We believe that much more work along these lines is urgently needed, and we predict that, if such work is not undertaken, practitioners in the field will start to quickly lose interest in the creation, use and demand for metadata and the capabilities it enables which would seriously reduce our progress to improve the effectiveness of learning.

More generally, metadata from other content that relates to the new content can be mined for relevant metadata – see section 3.4.

In addition, *templates* of reusable metadata can be created, where many of the relevant fields can be pre-filled. Often, instantiating the template will involve little more than simple selection between a small number of relevant values for a few remaining fields.

It is important to note that the" psychological" effect of presenting an automatically generated metadata instance and asking the end user to verify that this description is correct. This is a much less intimidating proposal than being presented with an empty form that includes a large number of empty text boxes to be filled in, as well as many long lists of values to be selected from.

3.4 Related content

Related content is another source of considerable amounts of metadata that is often overlooked. Indeed, further research is required on how metadata can propagate from one content object to a related one.

The following examples illustrate the kinds of propagation that can be considered:

- The language of a textual component will most probably be the language of the composite to which it belongs.
- If there is metadata available about the learning time of component learning objects, and about the way they are sequenced together in a composite, then it seems safe to suggest that the total learning time associated with the composite is at least the sum of the learning times over the shortest possible path, and at most the sum over the longest possible path.
- Content clustering techniques can be used to describe automatically the subject of a content component.

There are obvious parallels to this kind of metadata propagation and basic object-oriented modeling concepts like inheritance. A systematic analysis of how these concepts can be transposed from their original context to this kind of application is beyond the scope of this paper. We strongly encourage the research community to tackle this issue.

3.5 The author

The *author(s)* is often the source of additional metadata, as authors mostly produce content in one or a few languages only, or in one content domain, or for one kind of audience (say university level), etc.

Even for authors that produce more diverse content, the range of relevant values for many elements can be reduced significantly when characteristics of the author are taken into account.

Moreover, authors can create profiles that list metadata that are common to all or most of what they do: indeed, authors could create a few such profiles for the different kinds of objects they author. A simple example would list common metadata for home authoring and a different set of such metadata for professional authoring. In the latter case, further distinction could be made between content authored for specific projects, or courses, etc.

However, as with the introductory comments, we want to emphasize the need to make this process of having authors add metadata, be as easy, intuitive and automated as possible. Continuing the earlier theme of metadata propagation, the aforementioned profiles could be initialized and enriched with the metadata for content that was already authored in the same or similar contexts.

3.6 Feedback on effect

In the case of learning objects, we should be able to make use of information about how the learning object helped the user and organization to achieve the goal in an effective and efficient way.

This too can be kept simple, easy and unobtrusive. Note how this is done for example with the simple “thumbs up / thumbs down” buttons in the TiVo application described section 5.

Similarly, some web sites with technical support information include a simple feedback mechanism that enables the reader to indicate how useful and relevant the information was.

Of course, this becomes slightly more complex when we would actually try to measure the learning result and how it relates to the use of the object involved. In this context, deployment over massive numbers would enable us to filter out idiosyncratic effects that could otherwise compromise such data.

In fact, the very fact that authors decide to include a learning object in a larger context is a form of feedback that can be mined. Learning Object Repositories could include Amazon-like social recommending techniques to suggest relevant content to those searching the repositories.

4 Manual Metadata are Fine too!

One of the prevalent myths we would like to expose is that quality of metadata can only be provided by professionals. This is grossly over-estimated as many studies have shown that metadata “amateurs” often do quite well and there is great power in the collective knowledge of a community or group. The trick is to get them interested to provide metadata in the first place.

One interesting example is the metadata for music on CD’s provided by *Gracenote*. When you insert a music CD into your computer, you may have noticed that your music player connects to “CDDB” which stands for CD Data Base. This enables the display of all the metadata for the music on that CD -- the names of the songs, their length, artist name, etc. You may not have even noticed this happening and just assumed, as most do, that all this metadata is on the CD itself, but it is not and instead comes to your machine via the Internet from this CDDB. Neither the publishers nor the music companies provide any of these metadata. Rather, the metadata are managed by the Gracenote company, a profitable organization that runs a business model based on selling the related software to the music player software companies and online services. Built into their software (what shows up inside the music players and online music sites) is a “submit” button that taps into Gracenote’s database which contains detailed information about every track on essentially every CD produced just about anywhere in the world. About a million users per day (often several times a day) use it to find out about their tracks.

Those same users also contribute about 7000 submissions per day to the database by clicking on the "submit" button usually included with music playback applications that reference CDDB. Of those 7000 submissions, about 1500 make it into the database each day. Those 1500, a mix of new CDs and updates to already-cataloged CDs, have survived several thousand filters that weed out spurious submissions, automated voting logic to select the most likely accurate version from among near duplicates, plus human screening (done in the US, Japan and China) when needed. Within hours of a popular new CD's release, Gracenote receives between 10 and 100 submissions of information about it.

In addition to a great example of mass contribution of metadata, Gracenote is also an excellent example of the kind of business models we will see much more of in the future – those that combine the seemingly impossibility of an "Open Source" type of approach in the form of CDDB with a very sustainable economic or business model in licensing the technology to use this metadata.

Another, somewhat more playful example is the *ESP* game developed at Carnegie-Mellon[3], which takes a fun, novel look at how to gather metadata from several people at once - the idea being that if more people (well, ... 2) agree on a metadata term, then the quality should be OK. The point is that this game hides the fact that participants are doing something useful and focuses the user experience on scores and other games related criteria.

The importance of the community aspect is also illustrated by sites like Slashdot or kuro5hin, that rely on ratings of submissions (yes, those are metadata too!) to publish or reject what they receive. Typically, turnaround time is very low (less than a day) and the number of people involved in this kind of quality assurance scheme can be quite high. No surprise, the perceived value of such communities by those who participate in them is very high indeed,

Finally, while not always the case, blogs (web logs) can be considered to be quite elaborate and very structured metadata, typically about other people's documents. Indeed, even providing links to documents from others adds metadata to those documents. That is precisely the kind of metadata that Google exploits through its PageRanking algorithm!

5 Subjective Metadata from Multiple Sources

For those who are unfamiliar with TiVo and Personal Video Recording technology, it is worthy of a short explanation to show how it is already enabling an early example of personalization through pattern recognition of metadata generated unobtrusively throughout the user experience.

Simply put, a TiVo is a VCR that replaces the video tape with a big hard drive so that instead of storing the recorded TV or video signals on tape, it puts them onto a hard drive.

The increased capacity and ability to capture streamed video on the fly, adds benefits such as pausing "live" TV when the phone rings or when it's meal time. When you are ready, you hit the Play button and it picks up where you left off. It doesn't take long to then realize that you can simply skip over commercials or jump to any spot in the show, rewind and repeat something you missed the first time, etc. Continuing with somewhat incremental improvements, some other of the programming features stand out such as the ease of indicating what you want to record. In the case of most PVR's such as TiVo, you simply indicate the name or key words of a type of show or movie you like, and let the software use this to scan the daily updated programming guide to find and record shows that match no matter when or what channel they are on. What you end up with at this point is "myTV channel" consisting of just the shows that you would like to watch.

[3] http://www.espgame.org/

But the real magic and value that we would want for personalized learning, comes from two "must" buttons on the remote control. As you are watching any program, recorded or live, you have the option to press either a "thumbs up" (indicating good, I like this) or "thumbs down" (meaning I don't like this). To indicate your level of like or dislike, you can press each button once, twice or three times. As this information is collected into the memory of the machine, it gets better and better at knowing your individual preferences. Using this increasingly detailed understanding of your preferences, you begin to notice that programs have been automatically recorded onto what we can understand to be "meTV" and all without you asking for these to be recorded or even knowing that these shows existed. Yet when you take a look at them, (just before deleting what you probably thought was a mistake), you find that in fact these are some of the best shows you've watched in a long time.

This functionality is based on simple but powerful pattern recognition. Finding common patterns of the attributes of what you liked and didn't like, sifting through the thousands of attributes it has (metadata) about each show and also comparing your attributes and patterns with others just like it to determine with high precision, just what you would like to watch. Even when it makes a poor decision, if you note this with the thumbs down button, it is now a little bit "smarter" and will do a better job next time. This is an early but good practical and existing example of what "agent" and pattern recognition technology can do and why it will become so prevalent and valuable.

6 Future Work

In order to make methodological progress with the issues mentioned above, empirical data need to be analyzed about the actual use that is made of metadata, by different classes of end users (indexers, searchers, content developers, etc.). Such analysis could further our understanding of actual use cases. Useful approaches include log analysis of repositories, usability studies of metadata tools, analysis of the actual content of repositories, such as the kind of content, the actual metadata, the actual use of that content, the actual annotations by users who provide feedback on their use, etc. [9].

Another approach would be to analyze the differences and similarities between metadata authored by independent indexers. In principle, as there can be more than one metadata instance for a content object, this kind of analysis can be carried out a posteriori. In practice, it seems that, at this moment, this approach will have to be based on specifically set up experiments.[10,11]

Actually, we need to take this line of research much further. Our team at K.U. Leuven has started to work on information visualization approaches as a radically different way of enabling access to relevant learning objects [4]. Similarly, social recommending techniques may help to suggest appropriate resources at the right time. Newer technologies for content syndication, like RSS, could be applied in this context as well. The overall goal is to provide flexible access to advanced functionalities for end users, without putting any additional burden on their side.

7 Conclusion

In this paper, we argued that the potential of flexible and powerful functionalities, enabled by a ubiquitous metadata infrastructure, can only be realized if we work harder on gathering and exploiting these metadata in an unobtrusive way. We hope that this paper will help rally the community to expand and deepen the R&D and implementation in production environments of approaches that contribute to this vision.

References

1. Ariadne, 2004 (May 2, 2004); http://www.ariadne-eu.org/.
2. Duval, E., 2004, Learning technology standardization: making sense of it all, International Journal on Computer Science and Information Systems, 1(1):33-43; http://www.comsis.fon.bg.ac.yu/ComSISpdf/Volume01/InvitedPapers/ErikDuval.htm.
3. Duval, E., and Hodgins, W., 2003, A LOM research agenda, Proceedings of WWW2003 - Twelfth International World Wide Web Conference; http://www2003.org/cdrom/papers/alternate/P659/p659-duval.html
4. Klerkx, J., Duval, E., and Meire, M., 2004, Using Information Visualization for Accessing Learning Object Repositories, Proceedings of IV04 - 8th International Conference on Information Visualization.
5. ProLearn, 2004 (May 2, 2004); http://www.prolearn-project.org/.
6. Donald A. Norman, *The Invisible Computer: Why Good Products Can Fail, the Personal Computer Is So Complex, and Information Appliances Are the Solution*, Cambridge MA, MIT Press, 1998.
7. S3 Industry Report on "Making Sense of Standards & Specifications: A Guide for Decision Makers" published by The Masie Center. Full paper available at http://www.masie.com/standards/
8. Hodgins, W., Chapter on “Into the Future of meLearning: What if the impossible isn’t?!” in the book *Learning: Rants, Raves and Reflections (A Collection of Passionate Thoughts About the World of Learning,* edited by Elliott Masie and the Masie Center and to be published by: Pfeiffer in the summer of 2004.
9. Najjar, J., Ternier, S., Duval, E., (2003). The Actual Use of Metadata in ARIADNE: An Empirical Analysis. *ARIADNE 3rd Conférence, 2003.* Available at: http://www.cs.kuleuven.ac.be/~najjar/papers/WWW2003_najjar.pdf/.
10. Kabel, S., de Hoog, R., & Wielinga, B.J. (2003). Consistency in indexing learning objects: an empirical investigation. In: Duval, E., Hodgins, W., Rehak, D., & Robson, R. (Eds.), ED-MEDIA 2003, Proceedings Learning Objects 2003 Symposium: Lessons Learned Questions Asked, Honolulu, Hawaii, USA, Association for the Advancement of Computing in Education, Norfolk, USA, 26-31, ISBN: 380094-49-5. Proceedings available from: http://www.aace.org/conf/edmedia/LO2003Symposium.pdf
11. Elizabeth D. Liddy, Eileen Allen, Sarah Harwell, Susan Corieri, Ozgur Yilmazel, Necati Ercan Ozgencil, Anne Diekema, Nancy J. McCracken, Joanne Silverstein, Stuart A. Sutton: Automatic metadata generation & evaluation. SIGIR 2002: 401-402