

# OpenDocument Format Metadata Use Cases

## Introduction:

The following use cases were developed by the OpenDocument Format Metadata SC for use in specifying the scope of further work by the SC and for the purpose of deriving requirements for future additions to the OpenDocument Format standard.

The use cases are presented in alphabetical order. For each use case, there is an overview of the general problem area followed by one or more illustrations of the use case.

## Use Cases

### Accessibility Information

#### Overview

Not all users have equal access to all the content that can appear in an ODF document. Some questions of accessibility involve the use of metadata that can be used with or as an alternative to the content of an ODF document.

Handling of matters such as information about the document itself, how to read tables within tables, a box of text in a column of text, footnotes and communicating the navigation of the document per se, are all issues that touch upon accessibility. (The metadata SC has contacted the accessibility SC for guidance on these issues.)

#### Scenario 1

A user wishes to use reading software that cannot process images or other binary data embedded in an ODF document. Metadata should both signal the software of the presence of the binary data as well as carry alternative information concerning the content of the binary object.

#### Scenario 2

An author wishes to make more information available about a binary object in a hierarchical format that allows a user using reading software to access a very high level description such as is usually seen as a caption or more detailed information that would be available if the user were not using reading software. Charts, graphs, photographs, maps and similar materials often contain complex information that remains unavailable if only simple structures are available for alternative metadata.

### Asymmetric metadata

#### Overview

There are many situations where documents are interchanged during negotiations and each party to the negotiation wishes to preserve the metadata it has associated with the document but at the same time not reveal that metadata to other parties in the process. Such metadata might include workflow or

approval information, legal advice, comments on changes made by other parties, etc.

## **Scenario**

A settlement document for an accident case will contain provisions relating to the compensation to be paid by one side to the other. If the settlement document's metadata included a comment on a proposed settlement figure, such as "Don't go above \$15,000 unless party X is also released." that would be information that should not be shared with other parties to the negotiation. Yet, it is reasonable for one party to be able to include such metadata in the document for internal circulation.

## **Automatically generated metadata**

### **Overview**

Applications operates in environments that may contain metadata that can be automatically imported into such applications. User's in such environments, either as local operating systems or networks, may have metadata about themselves recorded in such systems. Metadata about the origin or changes to documents may also be held by such systems or may be automatically produced by such systems.

Automatic metadata generated in an environment would include Vcard, foaf, contact records, etc. Mapping of such metadata, both into and out of ODF would need to be supported.

### **Scenario 1**

A user has opened an application that conforms to the ODF standard in an environment that contains metadata about the user. Should the application accept metadata from the environment for automatic inclusion in the document instance? And if so, should it be included with or without notice to the user? Furthermore, should such metadata be limited to ODF defined metadata or should it be permitted to include custom metadata, in the sense that the metadata has been defined by the environment?

### **Scenario 2**

A user wishes to import foaf metadata into an ODF document. This requires a reversible mapping to and from the document.

### **Scenario 3**

A user edits a document in an editor that supports semantically rich metadata. That document is then edited by another user with a less capable editor. The first user expects the metadata they entered to be preserved, even if unrecognized by the less capable editor.

## **Bibliographies and Citations**

### **Overview**

Most textual document include references to content from elsewhere. That referenced content might be

quoted excerpts, data summaries, or paraphrased findings or conclusions. In fields where attribution of such referenced content is essential, such as law and academic research, citations and reference lists associate referenced document content with their source. And yet, formatted reference lists typically represent a subset of the source metadata, and may need to be reformatted for different audiences. In this sense, citations and reference list items can be understood as dynamic text fields whose content is generated from linked metadata descriptions.

It would therefore significantly enhance the possibility for user collaboration and application interoperability to have a standard metadata infrastructure. Likewise, bibliographic metadata is more complex than the simple document metadata commonly found in productivity applications, which is often just a series of key/values. Consider a simple example of a journal article, which involves relations between a document and a periodical, one or more people who author that document, and so forth. Beyond standardization, then, it is important to have a metadata approach that can support that sort of richer description.

### **Scenario 1**

Three users collaborate on a paper, one using an OpenDocument-compatible application and the other two users using MS Word.

As they write the paper and add citations, the citations and bibliography are automatically generated from the embedded metadata. Because the metadata is embedded, it's also portable. When the users pass the document around, the logic is always there so that the formatting can be regenerated. And because the metadata is based on a standard model, it would also facilitate interoperability between different third-party bibliographic applications.

When authors finish paper, they send it to a publisher, who can extract the metadata and make it available to search engines and journal providers. A standard metadata model also allows the publisher to regenerate the citations in a variety of standard styles (such as MLA, APA, Chicago).

## **Content Tagging**

### **Overview**

Users may wish to associate metadata with content at the level of OpenDocument markup. The OpenDocument markup that should allow the entry of metadata includes spans, paragraphs, figures, and tables. Metadata should also be associated with content on the basis of styles to be applied to that content.

### **Scenario 1**

A user wishes to enter metadata for a paragraph, such as its security level or to add copyright owner information to a figure.

### **Scenario 2**

In legal publishing (and presumably other domains) it is quite common to take an existing document (usually published legislation) and manually tag it with semantic information. One example would be

RDFa, which is used with XHTML.

For example, a paragraph or series of paragraphs may constitute a legal definition of a term. A span of text may actually be a cross-reference within the same legislation, a reference to case law, or an amendment to another act.

Tagging is often a precursor to transforming a document into a domain-specific format.

In the more general sense, activities such as indexing and cross-referencing can be considered content tagging and should probably use the same mechanism.

## **Conversion Metadata**

### **Overview**

Conversion from other document or data formats will often reflect choices made by users in particular business environments. Preserving the mapping to ODF from particular data sources enables the reliable exporting of that data back to its source format.

For example, MS Fields would require a mapping into ODF and in the reverse in order to have useful round tripping between MS and ODF formats. The same requirements exist for conversion from formats such as DocBook to ODF and back to DocBook.

### **Scenario 1**

A user wished to convert an OOXML document into ODF format for interchange and modification within an open software business environment. After modification, however, the document needs to be returned to its source in the original OOXML format. To insure fidelity both in terms of how the data will be viewed in ODF as well as the export back to OOXML, the recording of the source to ODF metadata is necessary.

## **Document as Web Service**

### **Overview**

We are comfortable with the notion of HTML pages as document-like objects accessible for display at a location on the Web. We are also comfortable with the notion of a Web address as representing a remotely invocable service that responds to queries.

But today we generally think of non-HTML text documents as statically downloadable objects; they neither display themselves nor respond to queries. (PDF documents are an exception. They do display themselves, with appropriate browser plugins.)

An ODF document associated with a web address should certainly, like HTML and PDF, behave as a self-displaying object (presumably via plugin support). But beyond this, an ODF document should also be able to respond to web queries about “what sort of thing (document) are you?”.

To behave in this way, an ODF document should appear under appropriate circumstances as a Web Service. The answer to “what sort of thing are you” could be almost any of the metadata categories outlined elsewhere in this document.

Thus, whatever form the internal schematization of the metadata markup, it should be done with an eye to easy reformulation as the data description component of a web service.

## **Scenario**

A repository of large (>100 MB) ODF documents is made available for download via the web. Prospective downloaders would prefer not to have to download the entire document to determine if it meets their requirements. They should be able to issue a query to the document, understood as a Web Service, requesting it to tell something about itself. The return of the query would be some or all of the metadata contained in the document. Or, it might include a portion of the content of the document, as defined by the metadata. For example, the document might be enabled to respond to a request to reveal the content of its “Table of Contents” section; the boundaries of this section would be defined by metadata.

## **Enhanced Search**

### **Overview**

The biggest barrier to effective searching is that most data is basically very dumb. A search for “job” on Google will return results that include “job” as in employment as well as “Job” as in the Book of Job.

Enormous strides have been made in algorithms and techniques for searching but they all are limited by the limitations of the data which they are called upon to search. Currently, strings in a text are just that, strings in a text which may be matched or not in a search. But, the authors of those strings meant something in particular when they entered those strings in a data resource. If such authors had the ability to declare what was meant by any string, search applications could take advantage of that information to return more precise search results.

### **Scenario 1**

A genetics researcher is writing a paper and wants to use a gene name that is common between the mouse and human genomes. Support could (and should) be provided to annotate individual words in a text. But it is a lot of trouble to mark each term in the document. If the researcher could declare a vocabulary for the document, that is what a particular word or words mean in the document, any search engine could interpret those words to have the specified meanings.

### **Scenario 2**

A researcher is writing a paper and chooses one out of 3 legitimate names for a gene. Under current metadata regimes, the researcher cannot specify the alternative names that may be used by searchers. This means that the choice of the wrong search term may result in not finding the article in question. Search engines can store some mappings of alternative terms, but authors are in the best position to create mappings of such alternative names. And, recording that information in metadata, makes the

resulting information resource more valuable to all search engines, not just those with the required mappings.

## **Extrinsic metadata**

### **Overview**

Current mechanisms for metadata presume that such metadata is embedded in and travels with a document instance. While that is one implementation strategy, it also leaves users open to the discovery of metadata that they would prefer that other users not see. “Scrubbing” the document prior to delivery is one option, but then the user loses the ability to use the metadata associated with the document, at least in the “scrubbed” version.

### **Scenario 1**

Senior partner in a law firm delivers a document to a client that they have approved and the lawyer bills for having authored the document. The client inspects the metadata and discovers that in fact an associate is the actual author of the document. But the law firm needs the metadata in order to track work assignments within the firm.

### **Scenario 2**

A contract has been reviewed for a client and attorney's have made comments on the legality or wisdom of certain portions of the contract using ODF metadata features. Normally these are privileged communications but the client has sent a copy of the document to another party to the negotiations.

### **Scenario 3**

A user has read only access to a document (either on a local network or the WWW) and they wish to annotate that document. When they view their annotations, the read only document is displayed to them with their annotations/commentary correctly displayed as part of the document.

## **Intellectual Property**

### **Overview**

Published documents of all kinds often include content from elsewhere: images, data, and so forth. This content typically has rights information associated with it. Yet currently managing such information is a manual task. An author or production editor must obtain the file(s), and separately manage the rights information. In turn, they must manually add such information to the published text in the form of captions with copyright information and so forth. This can be both tedious and error prone. Allowing such metadata to be attached to such content would allow for more automated solutions.

### **Scenario**

A government agency prepares a report that includes summary tables of data acquired from a third

party. The document author embeds the table data in the document, and captions—including copyright and source information—are automatically generated.

## **Scenario 2**

A student includes a Creative Commons-licensed photograph in their report. The license and attribution are automatically extracted from the image metadata by the application and appended to the image caption.

## **Metadata templates**

### **Overview**

When creating a document template, it should be possible to associate particular metadata with named styles in the document template, which could be used to trigger behavior during the editing session or at some later processing of the document.

### **Scenario 1**

A user wishes to associate a “no-spell-check” hint with a quote style to avoid spell-checking of quoted material.

### **Scenario 2**

A user wants to preserve data in an XML island in an ODF document, in a different namespace, for later processing.

### **Scenario 3**

A legal publisher works with outside authors to create a newsletter. They give the author a word processing template to use for formatting articles. The "hints" attached to various styles are used to transform the final document into a document instance valid for a particular XML schema required by the publisher's internal publishing tools.

## **Ontology Validation**

### **Overview**

As was discussed under Enhanced Search, users are in the best position to know what metadata they should associate with terms in their documents. Another aspect of the problem is when there may be two or more such meanings within a particular document.

### **Scenario**

A user has declared personal terminology metadata for their document in addition to a default terminology template used in their business environment. If terminology definitions differ between the

two metadata sets, the user should be presented with the option of selecting which definitions should be active for a particular instance in the document or even generally for this a document.

## **Realtime Collaborative Editing**

### **Overview**

The requirements for realtime collaborative editing can be found at: The main idea is expressed in [[http://en.wikipedia.org/wiki/Collaborative\\_real-time\\_editor](http://en.wikipedia.org/wiki/Collaborative_real-time_editor)].

From a metadata perspective, the question is how handle multiple instances of metadata for the same content (word, paragraph, page, document) for multiple authors.

One practical ramification of this may be that metadata will always need to be expressed as XML elements, not as XML attributes, since you cannot have multiple instances of the same attribute on the same element.

### **Scenario**

Users have a realtime collaborative environment that uses ODF as its format. While the current standard supports having multiple author metadata elements, for example, the user wish to record different, possibly inconsistent metadata on either ODF elements or spans in the ODF document.

## **Revision Metadata**

### **Overview**

The current ODF standard specifies markup for maintenance of a revision number for a document, the last person who modified the document but doesn't provide a standard way to record the revisions that were in fact made to a document. Revision history, stored as metadata, is important information about the current composition of a document.

### **Scenario**

A contract has been prepared by several users working in concert and a question has arisen as to who is responsible for certain language in the contract. With revision metadata capabilities, the responsibility for changes to the document are recorded as part of the metadata associated with the document.

## **Rich Semantic Metadata**

### **Overview**

Some disciplines such as health care and bioscience have controlled vocabularies that provide standard, sharable identifiers for objects in their domain. These controlled vocabularies range from simple lists of concept identifiers; to hierarchically organized taxonomies; to (in a few cases) elaborate and extensive

ontologies that incorporate substantial domain knowledge within the logical relationships asserted among the terms.

Users may wish to attach such identifiers to document text at a granular level (e.g. the level of a sentence, a phrase or even a single word) in order to provide computer-processable representations of individual assertions in the text. This rich semantic markup might provide a basis for enhanced document searching, automatic document classification, or even machine inferencing about facts represented in the document, based on the domain knowledge residing in the source ontology or in a separate knowledgebase.

### **Scenario 1**

A hospital's clinical information system uses word processing software for its clinical documents (e.g. patient progress notes, operative reports, admission and discharge summaries) that generates ODF formatted files. It embeds concept identifiers from the SNOMED CT (Systematized Nomenclature of Medicine/Clinical Terms) vocabulary into the document (possibly in an intermediate natural-language processing step), to allow its computerized clinical decision support system to recognize statements about the patient's active clinical problems, drug allergies and risk factors. When the patient's file is accessed by a physician, the decision support system scans the available documents, aggregates the SNOMED CT assertions, and subsequently applies medical rules to issue warnings e.g. if drugs are prescribed to which the patient has been asserted to be allergic.

### **Scenario 2**

The patient is referred to another institution. The patient's medical records documents are forwarded. The computer-processable assertions in the standard SNOMED formalism travel with the ODF records. No separate clinical database file needs to be transferred.

## **Security metadata**

### **Overview**

Users often have permissions to see only parts of documents. If a document is stored on a secure network server, metadata attached to portions of a document could be used by an application to simply not render those portions for a user without the required authorizations. If saving other than to the server is disabled, users with varying permissions can work on parts of a document they are authorized to view while the remainder of the document is concealed.

### **Scenario 1**

Classification officers or those charged with such responsibilities in military and governmental offices must often decide what parts of documents can be released and that varies according to a complex set of conditions. And those conditions can change. If metadata could be affixed to a document according to security levels (developed outside of ODF) that would fit into the current needs of such classification activities. (Military/governmental)

### **Scenario 2**

Commercial enterprises often have documents that may contain sensitive personnel, marketing or legal information, while portions of the document need to be processed by staff without the required permissions. Metadata based security for ODF would enable the construction of applications that can use ODF in its native format (no additional features required) to meet the security needs of commercial enterprises as well. (commercial)

### **Scenario 3**

Consumers may have similar issues but absent proper network and server management, will need different capabilities to secure portions of documents. But, the same security metadata could support applications that selectively encrypt portions of an ODF document (PCDATA). The encryption aspects are beyond the scope of ODF, but the availability of metadata security information would support the development of such applications. (consumers)

### **“Semantic Web”- Ready Documents**

The WWW Consortium (W3C) Semantic Web initiative envisions the web evolving into a distributed network of “resources” (anything identifiable by a uri), linked by “properties” (resources whose specific purpose is to represent semantically-defined binary relationships among other resources). The web-- viewed as network (actually, a directed graph) comprised of such relationships--can be seen as an enormous, distributed database. The W3C's RDF/RDFS standard formalizes the system.

The key representational trope that enables this vision is that uri's shall be the “common currency” for the identification of things (resources) of any kind.

“Resources” in the intended sense can certainly include documents, parts of documents, or metadata atoms contained within documents. They can also include authors of the documents (persons, represented as a uri), or objects referred to in the documents (things, represented as a uri).

It is certainly desirable that any scheme chosen for applying metadata markup to ODF document content should be “friendly” to the use of uri's as metadata identifiers, and generally that it should support the easy extraction of contained metadata into an RDF form.

Standard ways of making markup “RDF friendly” are under development in the W3C and include the GRDDL proposal (<http://www.w3.org/2004/01/rdxh/spec>) and the RDF/A proposal (<http://www.w3.org/2001/sw/BestPractices/HTML/2005-rdfa-syntax>). These proposals do NOT require that the markup itself be natively expressed as RDF, but merely that it be transformable into RDF.

### **Scenario**

See examples in the two W3C proposals mentioned above.

## **Workflow Management**

### **Overview**

Documents of all types move through approval, editing and other cycles in a business environment. If a document carried metadata about its status in some defined process, it could be automatically routed to the next person in the cycle or reports/reminders generated about the handling of documents in the workflow.

### **Scenario 1**

All purchase orders require approval and orders above a particular amount require approval by a departmental manager. The purchasing department places all approved orders and needs to easily separate approved purchase orders for those that await approval. Staff who have submitted purchase orders need to determine when (or perhaps if) particular requests have been approved.

### **Scenario 2**

Reports to regulatory bodies, such as the SEC, often require approval by appropriate company officials with specialized knowledge. The ability to track who has or has not approved a particular document, perhaps including when the document was approved, would provide company officials responsible for such reports with an audit trail in the document concerning such approvals.

### **Scenario 3**

A user wishes to annotate portions of a document with a “todo” tag that indicates various tasks that need to be performed relative to the content. Not all editing or other tasks have the same urgency so the user wishes to include a priority on such “tags.” In reviewing “todo” items, the user wishes to sort the items by date of entry, date of completion as well as by who completed any particular “todo” item.