Improving findability through multidimensional organization of content

Rupert Westenthaler, Andreas Gruber (Salzburg Research)

> *"Findability, as defined by Peter Morville (2005), is the ability of users to identify an appropriate website and navigate the pages of the site to discover and retrieve relevant information resources. Wurman (1996, p. 16) observes that "The ability to find something goes hand-in-hand with how well it's organized." But locating information resources in the digital environment depends on more than organization. As Morville notes, findability is an inherently interdisciplinary concept that integrates practices of design, engineering, and marketing: Findability encompasses not only issues of organization and representation - two central concerns in the construction of an effective information architecture - but also seeking behaviour, interaction design, branding, search engine optimization, and Web standards, to name but a few of the considerations that can affect findability. The focus on findability is on facilitating and enhancing the user's overall experience with an information resource."*[1]

Editor's Note (March, 1st, 2010): This article is an early work-in-progress document. The work is partly funded by the Interactive Knowledge Project[2].

Comments to the authors (Rupert Westenthaler, Andreas Gruber) are very welcome!

[1] http://en.wikipedia.org/w/index.php?title=Findability&oldid=330315978

[2] http://www.iks-project.eu/

Abstract

We specify and justify, a proposed novel annotation scheme for describing real world situations. The motivation for the scheme is that content is about situations and therefore, the description of situations is crucial for finding any content relating to such situations. The idea is inspired by Richard S. Wurman's LATCH concept of organizing information. The scheme can be used to annotate travel offers, public events, activity streams, discussion threads etc.

We identified time, location, participant, category and hierarchy as the most important dimensions and describe for each of the dimensions, how annotations can be exploited independently of specific domain models or ontologies. For the dimensions of an annotation, we describe briefly its main elements and provide initial specifications as XML-Schema and annotation examples.

In our next steps we plan to define rules that can be applied to specific dimensions and we describe requirements and features that would be needed to create semantic components to exploit the annotation model for presentation and direct user interaction.

Why do we propose a new annotation schema?

Semantic Search was a major trend in 2009 which received broad attention in academia, industry and from end users. However, underlying technologies and models are still somehow unclear: What kind of machinery is needed to support better search results and an intuitive interaction with the user? The answers range from pure statistical approaches, linguistic analysis to metadata driven approaches which themselves require knowledge representation languages such as RDF or OWL for modeling and generating inferences.

Yet, the most mature and at the same time successful annotation schemes seem to be rather simple, e.g. Dublin Core (DC), BibTeX, the ACM classification or domains such as News (NewsML). They have the following characteristics in common:a) the classification is almost common sense or very specific to domains with a high degree of mutual understanding between the community members and b) they are representing singular and well established artefacts or objects such as books, journals and scientific articles, or news items. Hence the annotation objects of these schemes are very homogeneous in the way they are represented: simple and single items of web content: one news article, one pdf, one book!
However, content (on the web) consists of much more:

- structured descriptions of products, services and events (travel descriptions with booking feature, product descriptions together with stores, description of events such as conferences);
- social streams of activity updates (webblogs, microblogging messages);
- centralized discussion threads in various forums;
- complex collaborative working environments (e.g. google apps);
- gaming environments;
- etc.

These kinds of content are very diverse and therefore, it is unlikely that one common annotation schema for everything will emerge soon. Our proposed scheme may provide the necessary foundations for a more widely usable annotation framework: it is intended to provide instant benefits to the user and is extendible to the needs of a specific domain.

Which environment would be needed?

For this proposal to work properly at web scale, several conditions must be fulfilled:

- the proposed schema needs to be adopted by at least three communities and their standards - the web community and its major players, the semantic web community (W3C) and the content management community (e.g. trough a semantic CMIS);
- semantic lifting engines are available to extract such metadata from unstructured web content items;
- mappings for other metadata standards need to be defined;
- web content editors need to be able to store content annotations in HTML; and
- content management systems as well as browsers would need "meaning editors".

Inline web resource annotation and conceptual resources

Within this article we aim at specifying this novel schema for metadata, which can be used for both

1. inline annotation of web content by using microformats[3], RDFa 1.1[4] and HTML5 Microdata[5] and at the same time,
2. representing independent content objects e.g. a user query regarding real world situations.

The main advantage of having annotations within HTML is the improved machine readability of published content. Simple agents (e.g. Operator[6] for the Firefox browser) can make use of these annotations and provide extended user interaction with web content, such as retrieving events information from a local calendar.

Storing such objects in content management systems independently from the content items would ensure that they can be treated in the same way as other content items, thus would benefit from being revisioned and annotated properly and could be published as conceptual resource items or for the web of data – as so-called "non-information resources"[7].

Multiple dimensions for annotation of content

R. S. Wurman, in his insightful book Information Anxiety2[8] presented the acronym LATCH as a "finite" way of information organization a decade ago and which have been been described as one future trend[9] in content management systems. Wurman defined the following five dimensions:

- LOCATION as "the natural form to choose when you are trying to examine and compare information that comes from diverse sources or locales. If you were examining an industry, for example, you might want to know how it is distributed around the world. Doctors use the different locations in the body as groupings to

[3] http://microformats.org/

[4] http://www.w3.org/2010/02/rdfa/ - should be a recommendation in April 2011

[5] http://www.whatwg.org/specs/web-apps/current-work/multipage/microdata.html#microdata - this is the editors working draft, no schedule available

[6] https://addons.mozilla.org/en-US/firefox/addon/4106

[7] http://www.w3.org/DesignIssues/HTTP-URI2

[8] Richard S Wurman, "The Business of Understanding," in *Information Anxiety 2*, 2000

[9] http://stephanecroisier.jahia.com/top-trends-for-cmswcm-in-2010

study medicine. (In China, doctors use mannequins in their offices so that patients can point to the particular location of their pain or problem.)";

- ALPHABET as a method "to organizing extraordinarily large bodies of information, such as words in a dictionary or names in a telephone directory. As most of us have already memorized the twenty-six letters of the alphabet, the organization of information by alphabet works when the audience or readership encompasses a broad spectrum of society that might not understand classification by another form such as category or location.",
- TIME as an "organizing principle for events that happen over fixed durations, such as conventions. Time has also been used creatively to organize a place, such as in the Day in the Life book series. It works with exhibitions, museums, and histories, be they of countries or companies. The designer Charles Eames created an exhibit on Thomas Jefferson and Benjamin Franklin that was done as a timeline, where the viewers could see who was doing what, when. Time is an easily understandable framework from which changes can be observed and comparisons made."
- CATEGORY pertains to the organization of goods. Retail stores are usually organized in this way by different types of merchandise, e.g. kitchenware in one department, clothing in another. Category can mean different models, different types, or even different questions to be answered, such as in a brochure that is divided into questions about a company. This mode lends itself well to organizing items of similar importance. Category is well reinforced by color as opposed to numbers, which have inherent value.
- HIERARCHY to organize items by magnitude from small to large, least expensive to most expensive, by order of importance, etc. It is the mode to use when you want to assign value or weight to the information, or when you want to use it to study something like an industry or company. Which department had the highest rate of absenteeism? Which had the least? What is the smallest company engaged in a certain business? What is the largest? Unlike category, magnitude can be illustrated with numbers or units.

Two more possible dimensions have been inspired by the keynote talk of Paolo Traverso[10] during the I-KNOW09 conference: He introduced a model where he used TIME, LOCATION, PEOPLE and COST as his four main dimensions. His main message was, that in a world of — at that time — 50+ thousands apps for the iPhone, such a relatively simple model would be very powerful, if one could use all four dimensions in each application and if one could combine these aspects. E.g., it would be very useful if a calendar does not only know that possible events overlap in time but also to introduce some knowledge about spatial location into these systems in order to then allow "simple" reasoning by combining distances and the time one needs to overcome these with possible entries in calendars. Today, one could easily schedule a meeting ending to 1p.m. in Rome, Italy and then attend a conference on the same day at 2p.m. in London, UK. Short of tele-transportation, you cannot be in both meetings!

We therefore considered the categories PEOPLE and COST as very important in addition to the five LATCH dimensions. PEOPLE is a new and independent dimension which we abstracted a little further and called it PARTICIPANT. The new dimension COST is considered to be a specific space of the "hierarchy" dimension.

[10] Paolo Traverso, Keynote at I-Know 2009, Towards a Future Internet of Services and Content, Graz, September 2009

Users' goals and means in supporting a search / retrieval process

In order to decide which dimensions are important for users in their organisation of information, we defined for each dimension to what extent each of them is representing goals or means. Then this would allow us to order by those dimensions that ranked highest in terms of goals.

- Location as two dimensional space is a good means for searching information. In addition, the location of things is also a very widely used scheme for queries. So this aspect is important both as a means and as a goal.
- Technologies based on language analysis and statistical algorithms ("Alphabet") can be used to implement well performing search engines and are well researched. Therefore, alphabet is a good *means* for providing search functionality. But no one searches for letters nor for words. E.g., by writing "Barcelona" into an search field you may be interested in the location of the city; the costs to get there; important persons, buildings etc ... but not in the word itself. This is the reason why the alphabet does not have a high value as a *goal*.
- Time is very similar to location because it defines a "natural" search space which can be used as a means to provide search functionality and at the same time time is very well understood by users who search for information.
- Information provided by a category system are a good way to implement powerful search infrastructures, but categories per se, often fail to qualify as a goal. Only in cases where a category and its internal structure is very well understood and accepted by users, can it also become a good goal. However as a means it is domain independent - hence the above classification according to a means/ends analysis.
- Structuring information spaces along value ranges of properties is a very good means to implement search functionalities. Whether it qualifies as a goal depends again, on how important, well understood and widely accepted such a structuring is.
- The economic value (or the Cost/Price) of something is a scheme that is very well understood and important for users. It can be represented by either categories (cheap, ... premium, €1-€2 ... €10-€100) or by hierarchy (e.g. sorting items by the price). Cost is an important goal in many application domains.
- Humans are social individuals and their relations to other people and specific things are of high importance to them. Therefore such relationships are also an important goal for searches. But it is questionable, if such relations are a good means for providing search capabilities.

Main dimension vs. means and goals	**Means**	**Goal**
Location	++	+++
Alphabet	+++	0
Time	++	+++
Category	++	+
Hierarchy	++	+
Costs	+	++
Participant	+	+++

Legend: 0: n.a. + weak ++ medium +++ support		

Table 1: How different dimensions represent means or goals for the search and retrieval process.

The most important result of this exercise is that we eliminate "Alphabet" as an annotation dimension, and order the dimensions by importance:

1. Time
2. Location
3. Participant
4. Category
5. Hierarchy

Exploiting annotations along their levels of measurement

Stevens[11] proposed that measurements can be defined with respect to four different types of scales. These are nominal, ordinal, interval and ratio scale. The following table from Wikipedia[12], describes for each scale type, the permissible statistics, scale transformations and the mathematical structures which are defined for these scales

Scale Type	Permissible Statistics	Admissible Scale Transformation	Mathematical structure
nominal (also denoted as categorical)	mode, chi square	One to One (equality (=))	standard set structure (unordered)
ordinal	median, percentile	Monotonic increasing (order (<))	totally ordered set
interval	mean, standard deviation, correlation, regression, analysis of variance	Positive linear (affine)	affine line
ratio	All statistics permitted for interval scales plus the following: geometric mean, harmonic mean, coefficient of variation, logarithms	Positive similarities (multiplication)	field

Table 2: What can calculated from data of within a specific scale type (copied from Wikipedia).

The following table shows examples, how values of the annotation dimensions can be situated onto the various scales of measurement. If the algorithms of a specific scale can be supported by a system in general, it would be easily possible to exploit these algorithms for any domain specific feature.

We marked those elements as bold, that seem to be most attractive and important for a semantic retrieval machine. For those elements, we need to specify the rules, that can be

[11] S. S. Stevens, “On the Theory of Scales of Measurement,” Science, vol. 103, 1946, S. 677-680.

[12] http://en.wikipedia.org/w/index.php?title=Level_of_measurement&oldid=344805520

applied in order to support the search process properly.

Dimensions and Scales of Measurement	**Nominal scale**	**Ordinal scale**	**Interval scale**	**Ratio scale**
Time	**Relative temporal locations** (soon, later, in the future ...) **Time points/ periods referred by name** (Stone age, World War 2 ...) Such time points/periods may be converted to interval scale by some service.	Time periods can also be on ordinal scale: "WW2 starts/ends later than the Bronze Age")	**Dates are on an interval scale, because there is no natural zero for time.** Dates and times can be applied to e.g. Allen's temporal calculus.	All date/time definitions which using the current time as natural zero. Note that this is only possible at query time (e.g. search for events in the next 5 days, all news about global warming in the last two days)
Location	**Relative locations defined by free text or controlled vocabularies** (e.g. "penalty box", "7-meter area", "evacuation area"). **Locations referred by name** (Paris, USA, Europe ...). Such locations can be converted to interval scale points/areas by some service.	Relative regions classified by some ordinal scale (e.g. the green, yellow or red zone, where green means safe and red means evacuation).	**Points/Areas defined by some coordinate systems** (e.g. Latitude/ Longitude/ Altitude). This assumes that Prime Meridian, Equator and Sea level are no natural null points for lat/ long/alt)	All Points/Areas/ Altitudes relative to the current location. (e.g. "X" is twice the distance than "Y", search for mountains twice as high as the mont blanc using the sea level as natural null)
Participant	**Participants are always on an nominal scale.** Each participant is a nominal entry and only identity checks can be used to distinguish			

	different participants			
Category	**Categories are always on an nominal scale.** When using Taxonomies the hierarchy can be used to deduce additional categories, but still each of this category represents an entry on an nominal scale.			
Hierarchy		**User ratings and reviews** (e.g. from * to *****; from "cheap" to "expensive")	Some absolute annotations are on scales without natural zero (e.g. temperature)	**Most hierarchy annotations come with absolute values** (e.g. costs, dimensions, wights, population numbers ...).

Table 3: Which dimensions can be used at higher scales than the nominal scale?

Specification

Main elements of the annotation object for a setting

The structure of the annotation object is as follows: Time, Location, Participant, Category and Hierarchy. An annotation object consists of five annotation elements, where the temporal annotation is mandatory, multiple participants, a location as well as categories and hierarchies are optional.

1. Time (Interval, Point) is mandatory and for each annotation a single value.
2. Location (Point, Shapes) is optional and for each annotation a single value.
3. Participant (Named Entities) is optional and categorized in the following (optional) sub-dimensions with multiple values.
    1. Agent
        1. Person
        2. Organisation
        3. Community
    2. Document
    3. Instrument
4. Category (Taxonomies) is optional with multiple values in different category spaces.

5. Hierarchy (Value ranges for units of measurement), multiple values in different hierarchy spaces.
    1. Cost
    2. Review

Please note: In this article, we use EBNF[13] in order communicate main ideas in a simple notation, NOT for a formal specification.

```
annotation = time , [location] , [participant] , [category] , [hierarchy];
```

A simple example for a content annotation: Announcement of a jazz concert

Content: "http://www.mypage.at/concert-mirabassi-salzburg.html", title: "Giovanni Mirabassi in Salzburg"
Date: "2010-04-01T20:00:00"
Location: "Salzburg"
Category.MusicGenre: "Jazz"
Hierarchy.Cost: "€50"
Participant.Agent.Person: "Giovanni Mirabassi"

A simple example for querying for content which follows this annotation scheme and retrieving the instances matching the described situation

Content: ? (to be queried)
Date: "now+"
Location: "Salzburg"
Category.MusicGenre: "Jazz"
Hierarchy.Cost: " >= €40"
Participant: ?

Common definitions for all annotations

A Symbol defines an unique annotation. If present, a domain-specific code can be used for identification of the symbol (e.g. geonames code in the example below).

```
symbol = ((label, [xml:lang]), [code]);
```

XML Schema for Symbol:

```
<element name="symbol" type="tns:symbolType"></element>

<complexType name="symbolType">
   <sequence>
```

[13] http://en.wikipedia.org/wiki/Extended_Backus%E2%80%93Naur_Form

```
        <element name="label" type="tns:labelType"
minOccurs="1" maxOccurs="unbounded"/>
    </sequence>
    <attribute name="code" type="anyURI" use="optional"></attribute>
</complexType>

<complexType name="labelType">
    <simpleContent>
        <extension base="string">
            <attribute ref="xml:lang" use="optional"></attribute>
        </extension>
    </simpleContent>
</complexType>
```

Examples:

- Berlin as defined by geonames.org

```
<symbol code="http://www.geonames.org/2950159">
  <label xml:lang="de">Berlin</label>
  <label xml:lang="en">Berlin</label>
</symbol>
```

- A simple natural language Symbol

```
<symbol>
 <label>Giovanni Mirabassi</label>
</symbol>
```

#The label and the language is used as symbol of code is not present. If no language is present, than the label is assumed to be valid for any language.
label = xsd:String;

#The identifier of the annotation. If defined the code is used as unique identifier for the annotation.
code = xsd:anyUri;

#The language of the label
xml:lang = xsd:String

#A symbol on an ordinal scale
ordinalSymbol = symbol, [compareCode]
TODOs: Schema + Example

#The code used to compare different symbols on an ordinal scale. If not present, the symbol it self shall be used.
compareCode = xsd:String;
TODOs: Schema + Example

Time definition

Time annotations can be made with either relative or absolute times. A relative time is just a symbol or may use a code. Absolute time annotations would be for a time point (a single date/time entry) or a time span by using start time and end-time or start-time and duration.

```
time = symbol | (timePoint | timeSpan) | "unknown");
timePoint = xsd:DateTime;
timeSpan = xsd:DateTime, (xsd:DateTime | xsd:Duration);
```

XML Schema for Time Annotations:

```
<element name="time" type="timeAnnotationType"></element>

<complexType name="timeAnnotationType">
  <sequence>
  <element ref="symbol" minOccurs="0"
maxOccurs="unbounded"></element>
  <element name="time" type="tns:timeType" minOccurs="0"
maxOccurs="1"></element>
  </sequence>
</complexType>

<complexType name="timeType">
   <attribute name="start" type="dateTime" use="required"></attribute>
  <attribute name="duration" type="duration" use="optional"></attribute>
  <attribute name="end" type="dateTime" use="optional"></attribute>
</complexType>
```

Examples:

- Time annotation referring to named time periods

```
<time>
<symbol>
   <label>World War II</label>
</symbol>
</time>
```

- Time annotation referring to a time point/period

```
<time>
  <timeInterval start="2010-03-02T10:30:00" duration="PT30M"
end="2010-03-02T11:00:00"/>
</time>
```

- Time annotation combining a symbol with an absolute definition of the time range

```
<time>
   <symbol>
      <label>"The 30s"</label>
   </symbol>

  <symbol>
    <label>World War II</label>
  </symbol>
   <timeInterval start="1939-09-01T00:00:00" end="1945-09-02T00:00:00"/>
</time>
```

TODOs:

- unknown time annotations are not yet defined
- usage of Date should be possible instead of DateTime

Participant definition

Participants are individuals (i.e. named entities), which participate in the situation one wants to describe. We suggest three possible types of participants. Agent is the set, which comprises of people, organization and community; documents would be e.g. content items, that are linked to the description, instrument can be used to describe necessary equipments.

```
participant = {symbol, agent | document | instrument | unknown};
agent = person | organization | community;
```

TODOs: Schema + Example

Location definition

Location annotations can be relative or absolute. Absolute locations can be expressed as string symbols or using the respective longitude and latitude values of a geo-point, -rectangle, -shape or -area.

```
location = symbol | (geoPoint | geoRectangle | geoShape | geoArea);
geoPoint = (wgs84_pos:lat,  wgs84_pos:long, [wgs84_pos:alt]);
geoShape = {geoPoint};
geoRectangle = (geoPoint, geoPoint)
geoArea = {geoPoint}
```

TODOs

- Schema + Example
- are there any standards for geo-position notations to be considered?)

Category definition

Category annotations can be made within multiple category spaces, which can be defined locally by the content management system or standard taxonomies available on the web e.g. the NewsML taxonomy.

category = {(categorySpace, symbol)};
categorySpace = symbol;

TODOs: Schema + Example

Hierarchy definition

Hierarchy annotations can be made in multiple hierarchy spaces, which can be defined locally by the content management system or declarations defined in standards. Each annotation consists of the space, the annotated value and the used measurement unit. Annotations of different spaces but using the same measurement unit can be combined by queries.

hierarchy = {(hierarchySpace, symbol, measurementUnit)}
hierarchySpace = symbol
measurementUnit = symbol

TODOs: Schema + Example

Roadmap for further work

v1.0.0 (April 2010)

- Add considerations w.r.t the Interactive Knowledge Stack layers interaction, presentation, rules, model, storage.

v0.3.0 (March 2010)

- Work out rules that are needed for a semantic engine (subsumption rules, grouping of labels in the same language, temporal calculus).

v0.2.0 (March, 2010)

- Finish XML-Schema for every element and the entire annotation object.

v0.1.0 (March 4, 2010)

- Purpose and the overall model for annotation.
- Initial ideas for the specification of the annotation object and its elements.
- Initial demo system: In the current state, readers can have a look at a brief demo system implemented using exhibit and faceted search: http://www.salzburgresearch.at/~agruber/iks-annotation/iks-metadata-exhibit.html

Acknowledgement

This work has been partly funded by the Interactive Knowledge Stack (IKS) research project (FP7 ICT Grant No. 231527).