GENERATING CONTENT-RELATED METADATA FOR DIGITAL MAPS OF BUILT HERITAGE

SEBASTIAN MATYAS AND CHRISTOPH SCHLIEDER
Laboratory for Semantic Information Processing
Otto-Friedrich-University Bamberg, Germany
Email: {sebastian.matyas, christoph.schlieder}@wiai.uni-bamberg.de

Long-term monitoring of the physical state of a building is of crucial importance in the preservation of built heritage. Monitoring results are typically documented by sequences of digital maps showing the evolution of the preservation state through temporal snapshots. Over a period of several years the collection of digital maps grows very large which is why users ask for semantic search methods to access the collection, i.e. for complex information retrieval queries of the type *concept@location*. This raises the problem of generating adequate content-related metadata supporting semantic search. We show how to use the method of thematic projection to automatically generate content-related metadata for digital maps used in built heritage preservation. Furthermore, we suggest using these metadata for approaching another problem of practical importance, namely data quality. Problems with data quality arise whenever maps are produced collaboratively by different observers. We describe a method that permits to retrieve sets of digital maps with potentially conflicting information. The method is based on an extension of thematic projection and *concept@location* queries allowing to handle temporal concepts and similarity measures.

1 Introduction

In the area of built heritage preservation, traditional paper-based workflows are rapidly evolving towards workflows based on digital documents, especially digital maps. The development could render all forms of paper documentation superfluous as predicted by Ogleby (2004) [7] although it is unlikely that paper maps will completely disappear before a satisfactory solution for the long-term preservation (> 100 years) of digital documents is found. In principle, architectural drawings of historical buildings can be produced using standard software such as computer-aided design tools (CAD) or, in the case of urban ensembles, geographic information systems (GIS). In practice, however, public or private organizations involved in the preservation of built heritage have rather specific software requirements which is why they use specialized software tools which offer functionality that goes beyond that offered by off-the-shelf products. An example of a tool specifically designed to support mobile users on TabletPCs is the Mobile Mapping System (MMS) which has been developed at the Laboratory for Semantic Information Processing of Bamberg University. Digital maps produced with the MMS are highly complex digital documents which associate spatial objects with thematic information. Spatial information is structured by building-specific partonomies and thematic information by domain-specific ontologies. Figure 1 shows an

example digital map produced with the MMS.

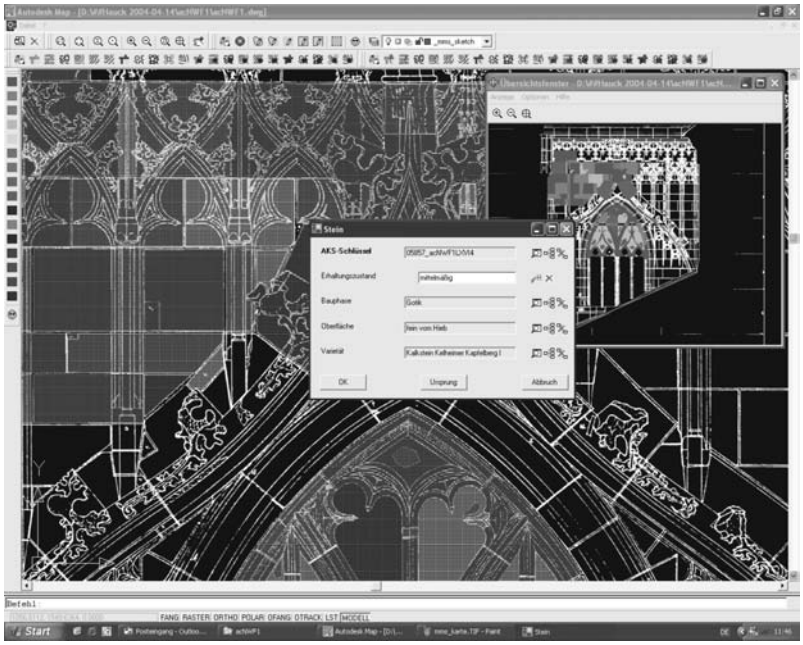

Figure 1. Mobile Mapping System

Digital maps are used in particular to document the results of long-term monitoring of the preservation state of a building. The evolution of the state is rendered by sequences of digital maps representing temporal snapshots. As a consequence, the number of digital maps from the monitoring process tends to grow very large as time proceeds. Without semantic search methods such a collection of digital maps would be very hard to access. Typically, the user wants to retrieve all maps that show a specific type of object at a specific place, e.g. "chemical erosion at the northwest facade". Semantic search should be able to handle generalizations and specializations in the domain ontology (e.g. chemical erosion IS-A stone damage) as well as in the building partonomy (e.g. northwest porch IS-PART-OF northwest façade). To support semantic search, the content of each digital map needs to be adequately described by metadata.

However, content-related metadata are fundamental not only for information retrieval. Another central problem in the context of long-term monitoring in built heritage is data quality. Monitoring constitutes a collaborative enterprise which involves many conservation scientists of different skill levels. It is not uncommon to find different interpretations of the same physical preservation state. Figure 2 shows digital maps produced by two different observers that were looking at exactly the same preservation state of an object but

recorded rather different spatial damage patterns (see Fitzner and Kownatzki (1990) [2] for details). It would be very helpful to detect sets of potentially conflicting digital maps based on content-related metadata.

Figure 2. Digital map produced by two different preservation scientists. Adopted from Fitzner and Kownatzki (1990)

The paper addresses the problem of automatically generating content-related metadata for digital maps produced during long-term monitoring of built heritage. For this purpose, we adapt the method of thematic projection from Schlieder and Vögele (2002) [8]and Vögele (2004) [12]. Furthermore, we suggest using content-related metadata for approaching the data quality problem. We describe a method that permits to retrieve sets of digital maps with potentially conflicting information. The method is based on modelling temporal concepts and by introducing similarity measures inspired by those used by Tversky (1977) [10].

The remainder of the paper is organized in the following way. Section 2 illustrates the method of thematic projection for generating content-related metadata for digital maps. The data quality problem is addressed in section 3. We describe how conflicting sets of digital maps can be found in the collections of digital maps. Section 4 discusses the approach in the context of related work and gives an outlook on future research issues.

2 Metadata in Built Heritage Preservation

Generally, the task of generating content-related metadata for complex documents is a very difficult one as it involves extracting the document's semantics. Automatically extracting a meaningful abstract from a text document, for instance, does not produce satisfactory results with current technology, because no complete computational account of natural language semantics has been given so far. Fortunately, the meaning of digital maps is much easier to describe than the meaning of a text because they are structured documents. The user provides valuable semantic information while producing the map. With the MMS, for instance, the user selects from a toolbar a drawing tool that is specific for a particular class of objects, e.g. stones. Through the drawing tool, every object created is automatically associated with information about its class. In addition, the user specifies properties and associations for most of the spatial objects using a domain ontology of his or her choice. The interpretation of properties and associations with respect to the domain ontology is transparent for the MMS and can be used to generate content-related metadata. In other words, the challenge of metadata generation consists not so much in extracting the semantics than in choosing what part of the semantic information to describe in the metadata.

A simple example document illustrates the content selection problem of metadata generation. Figure 3 shows a small detail of a map created with the MMS depicting a part of the cathedral of Passau, Germany. Graphical objects are arranged on three layers (often, many more layers are used). The background layer is formed by a raster image showing details of the architecture, in this case, a digitalized historical architectural drawing is used, in other cases one relies on photographs. On the base layer, the basic components of the construction, stones and joints, are represented. Objects of the base layer establish a spatial reference system with respect to which all other objects can be localized. The base layer objects are also called spatial objects. They are named using an identifier system specific to the building. Such systems have been developed by preservation scientist and can be extremely fine-grained. For instance, the identifier system created at the lodge of the cathedral of Passau, the AKS system, assigns a name to each individual stone of the building! Finally, the damage layer contains the damages as recorded by the creator of the map. Its objects are also called thematic objects. As simple as the example map is, it reflects quite well the central content of any map used in built heritage preservation: It relates thematic objects (in the example: damages) to spatial objects (in the example: stones).

The selection problem underlying the generation of content-related metadata consists in deciding which of these many relationships to describe in the metadata. According to Vckovski (1998) [11] several levels of abstractions for the resulting metadata are possible. In the case of maps of built heritage, this is an issue of spatial and thematic detail. On the one extreme we could store

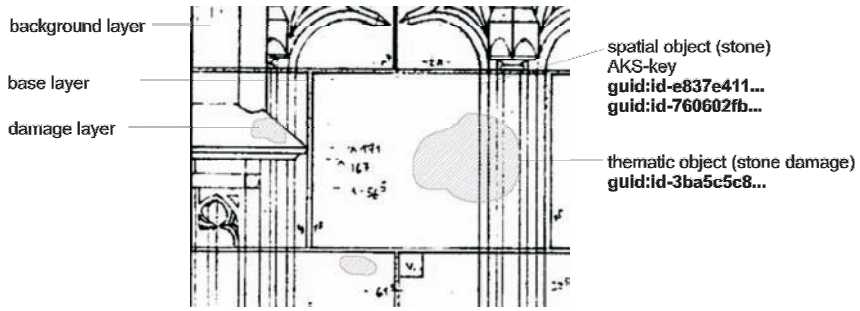

Figure 3. Digital map featuring three layers, one spatial object and three thematic objects (one stone damage "biogenous affection")

every bit of information about the thematic object in the metadata, including the geographic coordinates of every vertex. This would lead to duplicating the data and an enormous size of the metadata file. On the other extreme only highly abstract information could be stored like the names of the thematic objects drawn in the digital map. Although metadata file size is minimized that way, highly abstract content-related metadata are a poor guide in complex information retrieval queries. Typically they tend to lead to high recall and low precision result sets for queries.

Neither of the two extreme solutions is satisfactory. An intermediate approach has been proposed by Schlieder and Vögele (2002) [8], the method of thematic projection, in the context of metadata generation for geographic maps. It consists in (1) completely abstracting from all geometrical shape information, (2) in partly abstracting from information about position by using a system of objects that form a reference tessellation, (3) in providing the user the flexibility to choose how much thematic information to describe by metadata. Both abstractions are reasonable for built heritage maps, too. Users almost never need to formulate queries that involve a specific geometric shape, e.g. a damage having a certain polygonal form. A reference system is readily provided by the spatial objects that form the base map. Remember that each such object is uniquely named using the building identifier system. The thematic projection results from geometrically projecting the thematic objects of, for instance, the damage layer onto the spatial objects of the base layer. Flexibility is provided in so far as the user may choose the layers from which thematic projections are computed. To put it differently, the thematic projection of a digital map describes how thematic objects of the damage layer (and other thematic layers) overlap spatial objects of the base layers. It is this information that is encoded as content-related metadata. An extract of the metadata file created with the thematic projection will be presented in the following, generated for the digital map pictured in figure 3.

The resulting metadata file format is specified in the web ontology

language OWL [9]. It starts with a header listing document-related metadata according to the Dublin Core metadata standard [a]. Then the classes of the thematic objects and the spatial objects are recorded. In the example of figure 3 there is only one thematic class "stone damage" (guid:id-3ba5c5c8-...) for the thematic objects and the thematic class "spatial object (stone)" (guid:id-e837e411-...) for the spatial objects, both recorded as a *owl:Class*. In the object-oriented data model of the MMS, non-spatial data in form of attributes, like "AKS-Key" (guid:id-760602fb-...), are also stored as *owl:Classes* in the metadata file. Furthermore, associations between attribute classes and thematic classes are specified as a *owl:ObjectProberty* (guid:id-3d9222f4-...), here between "AKS-Key" and "spatial object (stone)". The semantics for attributes and associations are taken from the domain ontology. For the design of the domain ontology, which is individually done by the preservation scientist for every new building he is monitoring, no standard or top level ontology exists in built heritage. Therefore, the specification of the thematic classes, like "stone damage" or "stone" and also their attributes can range from simple textual annotations to complex objects.

```
<owl:Class rdf:about="urn:guid:id-e837e411-... ">
    <rdfs:label><![CDATA[spatial object (stone)]]></rdfs:label>
    <rdfs:subClassOf rdf:resource="&mmsdm;GlobalSpatialObject"/>
  </owl:Class>

    <owl:ObjectProperty rdf:about="urn:guid:id-3d9222f4-...">
      <rdfs:domain rdf:resource="urn:guid:id-e837e411-... "/>
      <rdfs:range rdf:resource="urn:guid:id-760602fb-... "/>
    </owl:ObjectProperty>
...
<owl:Class rdf:about="urn:guid:id-760602fb-... ">
    <rdfs:label><![CDATA[AKS-Key]></rdfs:label>
    <rdfs:subClassOf rdf:resource="&mmsdm;GlobalReference"/>
  </owl:Class>
...
<owl:Class rdf:about="urn:guid:id-3ba5c5c8-... ">
    <rdfs:label><![CDATA[Stone damage]]></rdfs:label>
    <rdfs:subClassOf rdf:resource="&mmsdm;ThematicObject"/>
  </owl:Class>
...
```

In the main part of the metadata file, the instances of the thematic classes, thematic objects appearing in the digital map, are recorded. In the metadata file for figure 3 instances of "AKS-Key" (guid:id-760602fb-...) and "stone damage" (guid:id-3ba5c5c8-...) are shown. Instances are either defined further as thematic projection concepts (*mmstp:Concepts*) or as locations (*mmstp:Location*), with the restriction that only the spatial objects of the base

[a] http://dublincore.org

layer can be locations. This is also the case for the spatial object (here guid:id-e837e411- ...) although they are not recorded as instances, because of their singleton attribute "AKS-Key".

```
<guid:id-760602fb-... rdf:about="urn:guid:id-ffe8001f-...">
    <rdfs:label><![CDATA[05875_acNOF1LVIII4]]></rdfs:label>
    <mmsdm:Reference rdf:resource="urn:guid:id-ffe8001f-..."/>
  </guid:id-760602fb-...>
...
<guid:id-3ba5c5c8-... rdf:about="urn:guid:id-03c6df54-...">
    <rdfs:label><![CDATA[affection, biogenous]]></rdfs:label>
  </guid:id-3ba5c5c8-...>
...
<mmstp:Concept rdf:about="#id-342dd23c-...">
    <mmstp:Represents>
      <guid:id-e837e411- .../>
    </mmstp:Represents>
  </mmstp:Concept>
...
<mmstp:Concept rdf:about="#id-918969b6-...">
    <mmstp:Represents>
      <guid:id-3ba5c5c8-.../>
    </mmstp:Represents>
  </mmstp:Concept>
```

In the last part of the metadata file the spatial relationship between the thematic projection concepts, i.e. thematic objects (#id-918969b6-...), and the locations, i.e. the spatial objects (guid:id-ffe8001f-...), are listed. Note that each attribute is represented as a class and therefore is stated to overlap the spatial object, too.

```
<mmstp:Location rdf:about="#id-294d06c3-...">
    <mmstp:Represents rdf:resource="urn:guid:id-ffe8001f-..."/>
      <mmstp:Overlapping rdf:resource="#id-918969b6-395e-..."/>
...
     </mmstp:Location>
</rdf:RDF>
```

3 Confliction Set Detection

With queries of the type *concept@location*, the user can retrieve maps that are likely to provide relevant information about the concept at the location specified in the query. A very similar type of query can be used to solve the data quality problem mentioned in the introduction. Figure 2 can be seen showing two members out of a possible large result set of a *concept@location* query. In other words, result sets may contain maps with conflicting evidence. Two questions immediately arise. How can sets of conflicting digital maps be

determined? Is there a simple way to aggregate the information contained in the conflicting maps?

To answer the first question, two situations that occur frequently in mapping of built heritage preservation must be considered. (1) Extensive use is made of base maps that are not clipped to the working area. Here the area that is actually being mapped is smaller than the base map. Because the working areas are not clearly separated redundant mappings of phenomena will occur. (2) New digital maps are generated by adding information, i.e. spatial and non-spatial data in form of thematic objects, to already existing digital maps, mainly base maps. A later retrieval for a specific area of interest would return all digital maps generated with this base map.

The intuitive answer to the second question would be to integrate all the information in the old digital maps into a new one. From a qualitative perspective this would be the optimum, because no information would be lost and errors would be corrected with information from independent sources. But considering the examples of digital maps shown in figure 4 this is not possible without interaction of a preservation scientist.

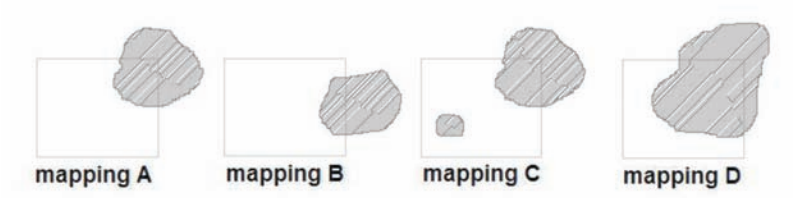

Figure 4. Possible variances of thematic objects in maps featuring the same spatial object

The problem with these digital maps is that only a preservation scientist can judge how to integrate their thematic objects into new ones. Because only domain knowledge tells whether the small damage depicted in the digital map C is one that was not present when map A was recorded or whether it was likely forgotten when producing map A. Similary, knowledge from preservation science is required to decide the following issues: Did the preservation scientist of the digital map A misplace the thematic object in the last preservation workflow, when digital map B was created? Has the damage probably even grown, or changed its position in a natural way in the digital map D from the time digital map A was drawn? Although, the integration of maps with conflicting information needs the involvement of a domain expert, the expert can be assisted by automatically retrieving confliction sets and by making them as small-sized as possible. Additionally the digital maps can be ordered for the evaluation of the preservation scientist, presenting the most conflicting pair first. The user could then select conflicting thematic objects together with an automatic method for integrating them in a digital map of improved

quality.

A prerequisite for determining conflict sets is to extend the *concept@location* query by a time concept. Ten year old base maps or damage maps are very likely not going to be relevant for finding conflicting evidence with today's maps. Almost certainly, some more recent map that derives from them will have exposed the conflict already. So they should not be integrated in the retrieved confliction set. The extended query would then have the changed syntax *concept@(location* $\wedge$ *timeframe)*. In the metadata file generated by the thematic projection for every thematic object the creation date has to be recorded, increasing the metadata file size only marginal.

To order the remaining digital maps an evaluating of the similarity of the digital maps has to be done. For the similarity measurement we adopted the contrast model of Tversky (1977) [10]. Tversky proposed to match features of compared entities and integrating them by the formula $S(A,B) = \theta f(A \wedge B) - \alpha f(A-B) - \beta f(B-A)$. Calculating the similarity for A to B, S(A,B) as a linear combination by matching the similar and distinguish features of A and B, respectively $f(A \wedge B)$ as well as $f(A-B)$ and $f(B-A)$. The terms θ, α and β are weighting the importance of these three parts for the similarity measurement. In most cases f is assumed to be additive. For a starting of future research we are weighting all three parts equally, setting θ, α and β to one.

Transferred to similarity between digital maps we have to extend the thematic projection about one further feature and consequently the metadata schema. During the thematic projection not only the condition of overlap of the thematic object and the reference entity will be recorded in the metadata, but also the similarity between these two objects. As features for the comparison we use the spatial coverage. The spatial coverage of the thematic object ($f(B-A)$), the spatial coverage of the reference entity ($f(A-B)$) and the spatial coverage they have both in common ($f(A \wedge B)$) are used as the similarity measures. By comparing the similarity measure S(A,B) we order the remaining digital maps, beginning with the pair of digital maps showing the biggest aberration, therefore presenting the preservation scientist the most conflicting pair first. This gives the resulting digital map the most qualitative increase, because the biggest errors are resolved with this first integration. Consequently we can extend the *concept@(location* $\wedge$ *timeframe)*query with the operand *similarity* for building the confliction set, defining the syntax as *concept@(location* $\wedge$ *timeframe)* $\wedge$ *similarity*. By using a threshold the preservation scientist can decrease the confliction set size once more because only pairs of digital maps are retrieved with a similarity lower than the defined threshold.

4 Discussion in the context of related work and future research

We described how to generate content-related metadata for the maps that are produced in built heritage preservation. The approach has been implemented in a mapping tool for preservation scientists, the Mobile Mapping System (MMS). It uses the method of thematic projection to produce metadata sufficient that is sufficiently informative to handle *concept@location* queries with high precision values. To achieve a high quality level for the digital maps created by different observers of different skill levels throughout the long-term monitoring of a monument two extension to thematic projection and the corresponding retrieval queries were proposed, namely the adoption of the time concept and the contrast model of Tversky (1977) [10].

Work on spatial metadata that is obviously related to ours are the spatial gazetteers used in the digital library research community. Gazetteers are place name lists that link names of geographic entities to geographic footprints and some sort of classification. The geographic footprint is either a simple geographic coordinate or a spatial reference on the basis of a regular, homogeneous grid. A very simple type of projection of a thematic layer to spatial reference entities is implied by gazetteers as described for example in Hill (2000) [5]. However, a simple form of spatial reference system is not appropriate for the complex compositional structure of historic buildings. Therefore, we used the objects named by the building identifier system as the reference system for the thematic projection of digital maps. It generates content-related metadata that is sufficient rich to enable complex information retrieval queries. This confirms results in Schlieder and Vögele (2002) [8], who used the thematic projection to generate content-related metadata for digital maps, generated by GIS.

Future work on the MMS will evaluate the proposed similarity measure and its value for the daily work of institutions involved in the preservation of built heritage. Further research is required to evaluate the best value combination for the weights in the contrast model that show the best results for ordering the confliction set. In this context, other similarity models should be analyzed too, such as alignment-based models (Goldstone 1994 [3]; Markman and Gentner 1993 [6]) or models based on the transformational distance (Hahn and Chater 1997 [4]). Although the *concept@location* retrieval possibilities in its current implementation appear to be sufficient in the practical usage, we are interested to integrate the full complexity of the 9-*intersection model* of Egenhofer and Franzosa (1991) [1] to further extend the retrieval possibilities for preservation scientist.

References

1. Egenhofer, M.J. and Franzosa (1991), *Point-set topological relations*, International Journal of Geographical Information Systems, 5(2), p.161-

174.
2. Fitzner, B. und Kownatzki (1990), *Bauwerkskartierung - Schadensaufnahme an Naturwerkstein*, Der freiberufliche Restaurator, 4, p. 25-40.
3. Goldstone, R. L. (1994). *Similarity, interactive activation, and mapping.* Journal of Experimental Psychology: Learning, Memory, and Cognition, 20, p. 3-28.
4. Hahn, U., and N. Chater. (1997). *Concepts and similarity.* In L. Lamberts and D. Shanks, Eds., Knowledge, Concepts, and Categories., Hove, UK: Psychology Press/MIT Press.
5. Hill, L. L. (2000). *Core elements of digital gazetteers: placenames, categories, and footprints.* In Borbinha, J. and Baker, T., editors, ECDL 2000, Research and Advanced Technology for Digital Libraries, Lisbon, Portugal, p. 280-290.
6. Markman, A. B., and D. Gentner. (1993). *Structural alignment during similarity comparisons.* Cognitive Psychology, 25, p. 431-467.
7. Ogleby C.L. (2004), *Heritage Documentation - The Next 20 Years*, XXth ISPRS Congress "Geo-Imagery Bridging Continents", Istanbul, Turkey, p. 850-855.
8. Schlieder, C. and Vögele, T., (2002), *Indexing and browsing digital maps with intelligent thumbnail*s, Symbosium on Geospatial Theory, Processing and Applications, Ottowa.
9. Smith, M., Welty, Ch., and McGuinness, D. (2004). *OWL Web Ontology Language Guide.* W3C Recommendation, World Wide Web Consortium (W3C), Februar 2004. http://www.w3.org/TR/owl-guide/.
10. Tversky, A. (1977), *Features of Similarity*, Psychological Review, 84(4), p. 317-352.
11. Vckovski, A. (1998), *Interoperable and Distributed Processing in GIS.* Taylor and Francis, London.
12. Vögele, T. (2004), *Spatial Information Retrieval with Place Names*, PHD. thesis, University Bremen.