D.6.3 – Third Year Report
WP6 – Creating the 3D Collection Item

Version 4.1 *(final)*

30 November 2011

Grant Agreement number:	**231809**
Project acronym:	**3D-COFORM**
Project title:	**Tools and Expertise for 3D Collection Formation**
Funding Scheme:	**FP7**
Project co-ordinator name, Title and Organisation:	**Prof David Arnold, University of Brighton**
Tel:	**+44 1273 642400**
Fax:	**+44 1273 642160**
E-mail:	**D.Arnold@brighton.ac.uk**
Project website address:	**www.3d-coform.eu**

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231809.

Author: **Achille Felicetti**

PIN, University of Florence, Italy

Contributing partner organizations: **University of Brighton (UoB)**

Foundation for Research and Technology, Hellas (FORTH)

Consiglio Nazionale Delle Ricerche – ISTI (CNR-ISTI)

Fraunhofer Gesellschaft Zur Forderung Der Angewandten Forschung E.V. (FhG-IGD)

Table of Contents

Table of Contents **3**

1 Executive Summary **5**

2 Introduction and Objectives **6**

3 T6.1 – Processing Tools for Metadata **7**

3.1 Multilingual Terminology Manager 7

3.1.1 Work performed 7

3.1.2 Deviation from work plan 8

3.1.3 Plans for the next period 9

3.2 Multilingual Support Library 9

3.2.1 Work planned 9

3.2.2 Work performed 9

3.2.3 Deviation from work plan 11

3.2.4 Plans for next period 12

3.3 Metadata Extraction tool 12

3.3.1 Work planned 12

3.3.2 Work performed 12

3.3.3 Deviation from work plan 13

3.3.4 Plans for the next period 13

3.4 Matching and DB Exporting Tool 13

3.4.1 Work Planned 13

3.4.2 Work performed 13

4 T6.2 Annotation and smart tagging tool **15**

4.1 Work planned 15

4.2 Work performed 15

4.3 Deviation from work plan 17

4.4 Plans for the next period 17

5 T6.3 Semantic propagation **18**

6 T6.4 Co-referencing resolution tool **19**

6.1 Work planned 19
6.2 Work performed 19
6.3 Deviation from work plan 19
6.4 Plans for the next period 19

7 Publications from WP6 20

8 References 21

Appendix A - Mappings to the SKOS model 1

8.1 Mapping ISO 2788 to SKOS model 1
8.2 Mapping AAT to SKOS model 1
8.2.1 Multiple languages 1
8.2.2 AAT's objects 2
8.2.3 AAT's concept example in SKOS data model 3
8.3 Mapping WebTMS xml format to SKOS model 4

Appendix B - Querying the MR vocabularies 5

1 Executive Summary

This deliverable describes the activities carried out during the third year of the 3D-COFORM project within Work Package 6 (WP6) by the different partners and describes the results achieved by this work package.

2 Introduction and Objectives

Work Package 6 includes the development of a toolset for the creation, management and integration of text metadata related to cultural digital objects and graphic processing. The toolset provides: a unified multilingual framework to support the various multilingual requirements of the project; technologies to enable scholars and Cultural Heritage users to generate metadata from legacy information to be shared in the semantic framework of the 3D-COFORM project (Task 6.1); annotation and smart tagging tools to create new metadata on top of digital objects and texts (Task 6.2); technologies for the preservation, the reusability and the consistent propagation of metadata alongside different digital objects (Task 6.3); tools for the management and the resolution of co-references issues in the digital object metadata (Task 6.4).

We have performed many tests on the various tools in different contexts to enhance their interoperability. Goals and achievements of WP6 for the third year are discussed in detail in the following sections.

3 T6.1 – Processing Tools for Metadata

3.1 Multilingual Terminology Manager

3.1.1 Work performed

The purpose of the Multilingual Terminology Manager is to handle the vocabularies and types that are used in the 3D-COFORM Repository. We defined a basic set of specific vocabularies, quite small (but expandable) and potentially multilingual.

Currently the RI includes the following vocabularies:

- Annotation Types
- Device Model Types
- Device Types
- Dimension Types
- Event Types
- Information Object Types
- Material Types
- Mime Types
- Physical Object Types
- Process Types
- Role Types
- Shape Types
- Software Types
- Software Version Types
- Type Types
- Unit Types

The Simple Knowledge Organization System (SKOS) [1] is the model used for the representation of the above vocabularies. SKOS, developed within the W3C framework, is a family of formal languages designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled vocabulary.

SKOS is built upon RDF and RDFS, and its main objective is to enable easy publication of controlled structured vocabularies for the Semantic Web. It is integrated with CIDOC-CRM through the E55 Type Entity.

The AAT thesaurus [2] was chosen by the 3D-COFORM consortium as the primary source for the definition of the types of all the physical objects that are handled by the 3D-COFORM Repository Infrastructure. These include the archaeological artefacts and the archaeological sites.

Finally, the implementation of the Multilingual Terminology Manager was based on the WebTMS application, a web-application for multidisciplinary bilingual thesaurus presentation and management that is based on ISO 2788 [3] and uses a simple proprietary XML-format for importing and exporting data [4].

The possibilities offered at the thesauri management level, cover a wide range of desired functions ranging from conservation and expansion of thesauri (functions of addition, modification and deletion of terms, hierarchies and facets, mass import / export of terms, creation of thesauri, saving and restoring backups etc.) to integration of multiple presentation methods (alphabetical, systematic graphical, hierarchical presentation) and access (alternative forms of navigation, support of complex search criteria, scalability search etc.)

The purpose of the system and of the underlying modeling is to cover all the needs that derive during the administration of thesauri by accelerating and facilitating the procedures necessary for their conservation (maintenance) according to the consistency checks specified by the ISO and ELOT standards.

In order to create our SKOS data model, we have produced a mapping between ISO 2788, AAT thesaurus and the SKOS model. Some elements had an exact match. For those elements for which SKOS did not have an exact match, we used some of the properties of SKOS vocabulary to create conceptual elements that correspond to those elements/notions of AAT thesaurus.

We have also implemented a library of methods for converting thesaurus data from one format to the other, based of the above mapping. With this library we have converted the AAT thesaurus data from AAT XML proprietary format to SKOS RDF-format. We have also converted the resulting SKOS RDF-formatted data to the WebTMS XML proprietary format and loaded the AAT thesaurus to the WebTMS application. Through the WebTMS application, the user can create the desired 3D-COFORM proprietary vocabularies, for example the vocabulary for the Physical Object Types.

For the needs of the Multilingual Terminology Manager we have defined three mappings to the SKOS model which are presented in Appendix A and a set of basic SPARQL queries for accessing the MR vocabularies presented in Appendix B.

3.1.2 Deviation from work plan

No deviation from the work plan.

3.1.3 Plans for the next period

We will extend the WebTMS application to support multilinguality.

3.2 Multilingual Support Library

3.2.1 Work planned

The key activities planned for Year 3 were: development of QT bindings for the MLSL to support integration into the IVB shell; further development of internal MLSL functionality, including text-search, networked thesauri and linguistic information management, and extension of linguistic coverage. We aimed to deliver the final MLSL release in Month 36.

3.2.2 Work performed

3.2.2.1 MLSL binding and IVB integration

Further functionality was added to the existing AnnoMADml demonstrator system, developed in Years 1 and 2. This system can provide a context for interface development very similar to the IVB context. This approach has the benefit of decoupling the MLSL development from any specific scenario (e.g. IVB), generating a modular design of the MLSL as a component of the 3D-COFORM framework. These developments should transfer to the real IVB with minimal additional effort.

3.2.2.2 MLSL functionality development

During this period we have further extended the facilities developed last year, and described in D6.2 – Second Yearly Report WP6, to support 'Active Resource Bundles', which allow much greater flexibility of linguistic control in localised text strings. In particular, we have added basic morphological support for phenomena such as mass and count nouns, singulars and plurals and verb tenses and irregular verb forms. We have also extended the initial summary generation resources to support dynamic captioning – the generation of brief descriptive text strings for example as pop-up 'tool tips' on artefacts displayed in browsing and viewing interfaces.

We have also developed more sophisticated tools for user-specification of metadata. This includes an improved thesaurus access interface, based on the SKOS thesaurus representation standard and with the capacity to access SKOS-based thesaurus resources (see section 4.1 above and also the discussion of external resources in D7.3), and new support for user-specification of multilingual terms (translations of user-supplied labels etc.).

3.2.2.3 Linguistic information management

A key decision made and implemented during this period was the way in which multilingual information is represented and stored in the metadata repository (MR). Metadata is stored using RDF triples representing relationships between entities (artefact, agents, locations etc.) These representations have text string descriptions associated with them, but these are often not suitable for presenting metadata information to the user, for example in a browser interface. The main reasons for this are:

1. The text strings may be (or contain) technical components, not just user-friendly text (for example, they may be URIs or ontology descriptors such as 'P3F.has_note')

2. Even when user-friendly text strings are used, they may not be provided for the required language, and although the user interfaces can be designed to standardise the language used for known fixed terms (such as thesaurus entries), it is not feasible to require all user-generated content in the MR to be specified in every supported language.

3. Simple text strings do not in general contain sufficient information to support use in more advanced linguistic contexts, especially for languages with non-trivial morphology. For example, they will not include different inflectional classes or irregular forms for verbs, or mass/count information for nouns. Such information is of significant benefit to tasks such as caption generation and natural language summaries.

To address these issues, the MLSL employs an architecture which assumes these text strings are 'conceptual' rather than linguistic, and only include language-like tokens on an 'as available' basis, without any guarantee of coverage or suitability for a particular user-interface requirement. All the linguistic information presented to the user is stored in the MR independently, with the proviso that, if such information is absent, the system can fall back to using the conceptual strings where they are available. As well as supporting multilingual requirements more flexibly, this approach results in a clean separation between the core conceptual knowledge in the MR, which has considerable value as a resource in itself and may be processed by systems with no interest in linguistics, and the more ephemeral linguistic knowledge, which is primarily present to support user interface localisation.

However, requiring that conceptual and linguistic knowledge should be separate does not determine absolutely how linguistic knowledge should be represented, and a number of options are possible. First, the XML standard underlying the MR representation framework provides limited support for multilinguality: string literals may be annotated with language tags which can then be taken account of by applications. However, this approach still only provides support for simple strings, not more complex linguistic information, and conflates the conceptual and linguistic information into a single system – essentially a multilingual conceptual system, potentially requiring all applications to be language-aware.

A second approach would be to extend the ontological framework to support linguistic information directly in new RDF triples. Such information could safely be ignored by applications without user interfaces, but would be available if required. This is a more flexible solution, but relatively heavyweight,

in terms of the additional ontological machinery and RDF triples that would be required to encode linguistic information associated with each conceptual entity. It is also structurally rather different from the standard libraries used for application localisation ('locale' chains containing keyword/value pairs) so considerable mapping between representations would be required.

A third approach is to encode complex linguistic information relating to a conceptual entity within a single string literal. This is a much smaller change to the ontological schema, encapsulates linguistic knowledge in a single RDF triple that applications can ignore if they wish, and allows the internal representation of linguistic knowledge to be aligned with the localisation model used by user interface libraries.

It is this third option which has been implemented for use in 3D-COFORM. The XML standard allows us to create 'typed literals', which are strings with an associated 'datatype' tag which indicates how the string should be interpreted (such strings may contain markup, for example, or program code). We have created a new datatype tag to represent strings containing MLSL linguistic information, in the same format as the property files used by the MLSL localisation mechanism. The MLSL library code reads such a string from the MR, compiles it into a lightweight property table and integrates it into the core property tables used by an application to localise its interface. The net effect is that it is possible to store arbitrarily complex, multilingual linguistic information in such strings on a per-concept basis, created, managed and interpreted by MLSL functions and stored in the MR, but completely encapsulated within the MLSL subsystem – other components of an application or users of the MR need know nothing about it.

3.2.2.4 MLSL v3.0 beta release

This year included development towards a new release of the MLSL scheduled for Month 36. In the original work plan this was intended to be the final release of the library, however, delays in the integration process mean that some further work is now anticipated in Year 4. Nevertheless, we are in the final stages of delivery of an interim release, MLSL v3.0 beta, for Month 36 as originally proposed. This release includes a substantial refactoring and repackaging of the library code in its ActionScript form to provide a cleaner and better documented interface, as well as the functional enhancements described above.

3.2.3 Deviation from work plan

Most of the planned activities have progressed, to at least some extent, and it has been possible to hold over some effort resources to complete this work in the first part of Year 4. The integration of the MLSL development into the IVB (with the RI) will then be conducted in Year 4.

3.2.4 Plans for next period

Most of the core functionality of the MLSL is present in the v3.0 beta release. The main requirements to deliver the final MLSL involve completing the integration into the IVB (discussed in more detail in D7.3) and making final refinements to MLSL data resources. We aim to complete this work and deliver the final MLSL by Month 42.

3.3 Metadata Extraction tool

3.3.1 Work planned

The main aim for this period was to improve and evaluate performance of the generic metadata extraction tool developed in Year 2.

3.3.2 Work performed

At the end of Year 2 we delivered an alpha release of the metadata extraction tool as a freestanding version of AnnoMADml, and noted that the technical requirements for installing this system were rather complex. During Year 3, we came to the conclusion that this system was in fact too complex (technically) for viable delivery as part of the main 3D-COFORM suite, and decided that we needed to review its overall design, with the intention of making it more modular, using more lightweight components. We compared our existing implementation with an approach based around the UIMA protocol ("Unstructured Information Management Architecture" – see http://uima.apache.org) a recent standard for modular architectures for text processing, originally developed by IBM. We concluded that the lower infrastructure overhead and industry-standard status of UIMA made it a better choice both for flexible development and final deployment of the tool chain to support metadata extraction. We therefore elected to refactor and repackage our existing extraction chain using a UIMA-compliant approach, resulting in a simpler installation process.

This change also prompted a review of the 'off-the-shelf' components in our extraction pipeline, and we adopted a combination of Apache openNLP (http://incubator.apache.org/opennlp/) preprocessing modules combined with the ENJU parser (http://www-tsujii.is.s.u-tokyo.ac.jp/enju/). These feed into a new bespoke relation extraction tool which we have implemented to interface to AnnoMADml. One incidental benefit of these changes is that the metadata extractor no longer has any dependence on Linux, and will in principle run on Microsoft Windows, in accordance with the project policy on delivery platforms.

This additional work on the tool chain has meant that our work on improving and evaluating performance was delayed. However, the new architecture also has the benefit of increased modularization of the statistical models which can be tuned to the domain, especially in the preprocessing steps. We have begun to explore the benefits of tuning these various components using

Louvre website data and AAT thesaurus data (described in more detail in D7.3) and aim to produce useful results in the early part of Year 4. We have also integrated the two versions of AnnoMADml back into a single system, incorporating both the metadata extractor and the MLSL extensions discussed above.

3.3.3 Deviation from work plan

We have experienced a minor delay in performance improvement and evaluation, partly due to unanticipated but very valuable work on the infrastructure discussed above, partly due to staff availability. We do not anticipate any problem catching up from this delay in the early part of Year 4.

3.3.4 Plans for the next period

We aim to complete our work on performance tuning and evaluation by Month 42.

3.4 Matching and DB Exporting Tool

3.4.1 Work Planned

For Year 3 we planned the improvement of the Metadata Extraction tool to create a stable release and the evaluation of the metadata performance of the Metadata Extraction tool. We also planned the implementation of the transfer mechanism to store the RDF information extracted from legacy databases into the Repository Infrastructure and the integration between the two tools, in order to produce a complete mapping platform.

3.4.2 Work performed

During the activities of the third year under WP6 we have improved the mapping process to create a stable version of the mapping interface. We have developed a web-based user interface of the Matching Tool providing a rich set of flexible functions for conceptual matching of schemas (see Figure 1). We have also written the scripts to perform the operations required by the DB Exporting Tool to export content from databases into RDF in accordance with the conceptual matching provided by the new Matching interface. The new scripts are based on the D2RServer model and take advantage of its rich syntax to implement the mapping on data and to filter the output of the extraction engine using domain-specific knowledge of the kind of information we are trying to extract.

The whole mapping/exporting pipeline has been tested on a dataset provided by the Victoria & Albert Museum (VAM). As a preliminary activity to the tests, we performed a data transformation definition

based on a conceptual mapping of the VAM dataset to CIDOC-CRM in order to establish precise correspondences between the elements of their database (MySQL) and various CIDOC-CRM entities and to build a rich semantic network used to populate the Metadata Repository.

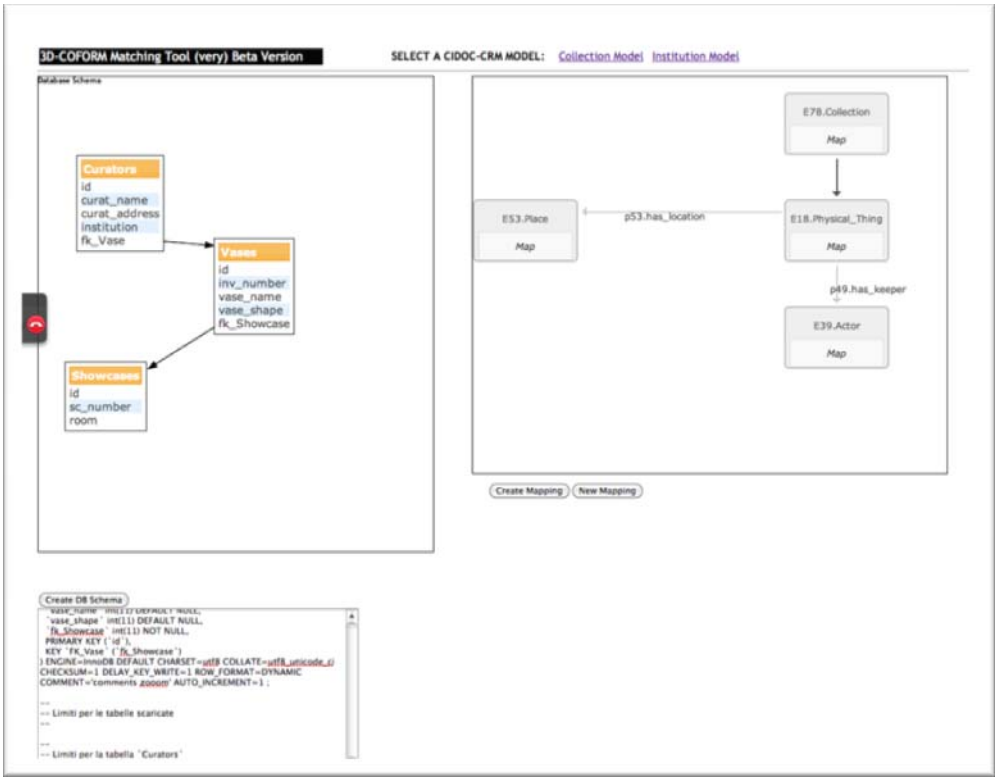

Figure 1: Matching Tool Web Interface.

4 T6.2 Annotation and smart tagging tool

4.1 Work planned

The Annotation module will be represented by several user interfaces using the annotation templates. Each one of these will support a different annotation scenario. The TriG-File Generator will support all types of annotation scenarios which will be provided by the annotation templates and it will generate out of the collected data from the user interface a RI compatible TriG-File. The Annotation Tool will provide a visual representation of an annotation. The development plan for this tool in Year 3 included also the possibility to deal with various thesauri.

4.2 Work performed

As described extensively in D3.3, we have implemented the MR internal ingest, update and delete functions in order to support the needs of the integration of the Annotation Module with the RI. The annotation metadata differ from the rest of the metadata because of the use of the concept of named-graphs in the semantic network.

The annotation tool as part of the IVB (WP7) was tested by several CH professionals. The most important feedback from these exercises was that it was not intuitive to change the window of the tool for creating the annotation after the geometric definition. Thus, we developed a new interface with new interaction concepts, which allow the user to create the geometric definition (in other words an area) of the model and by dragging and dropping, to indicate the areas to be annotated. Figure 2 shows the new annotating interface in the same window of the viewing interface, enabling a more intuitive annotation process.

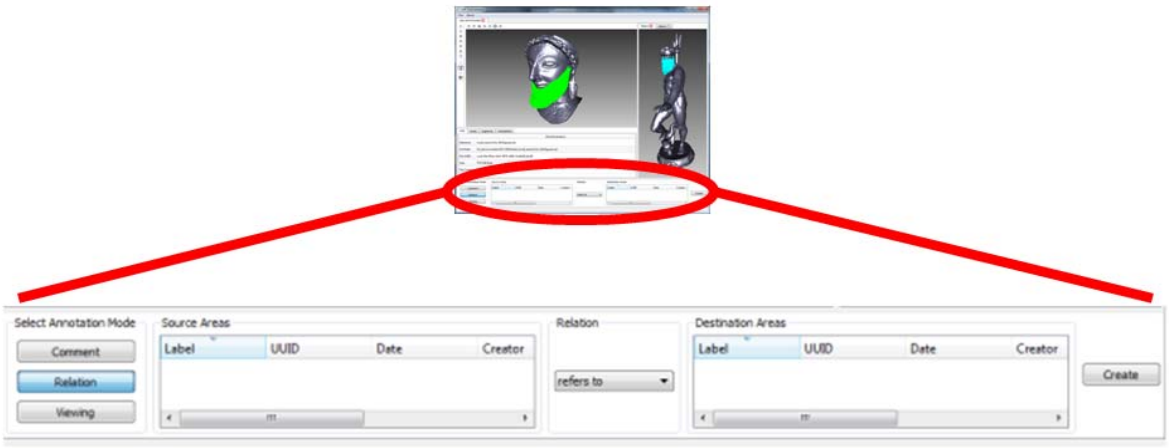

Figure 2: Integrated annotating interface with viewing interface.

The new interface allows the user to have a visual feedback of the geometric definition (area), while making an annotation. Thus, the user can clearly identify the areas for which the annotation is created. The annotation interface provides three modes: i) comment, ii) relation, and iii) viewing.

- The comment mode enables the creation of a comment associated to one or more areas.
- The relation mode permits the creation of a relation between multiple areas.
- The viewing mode provides an interface for visualizing either a single annotation or the associated annotations to an area; in both cases by dragging and dropping an annotation / area from the sheet of the viewer (information about the viewer is found in the deliverable D.7.3, Task 7.1).

A consistent set of information is always presented independently from the selected mode, thus label, UUID, date, and creator are always visible, nevertheless the user can also select different fields or add additional fields, and it is also possible to sort each field.

The concepts behind the IVB, including the annotation component were published at VAST 2011 ([T6.2.1]). Additionally, the first state-of-the-art in 3D annotation was submitted to Eurographics STAR 2012 ([T6.2.2]) and a short version was presented in the public 3D-COFORM STAR Workshop at VAST2011. Figure 3 depicts the building blocks of the annotation process, which are covered in the STAR. The solution proposed in 3D-COFORM is the first integrated solution covering all the building blocks. Nevertheless, in the community there is no notion for representing annotations, thus there are interoperability issues, which should be addressed, in order to achieve integration beyond individual initiatives.

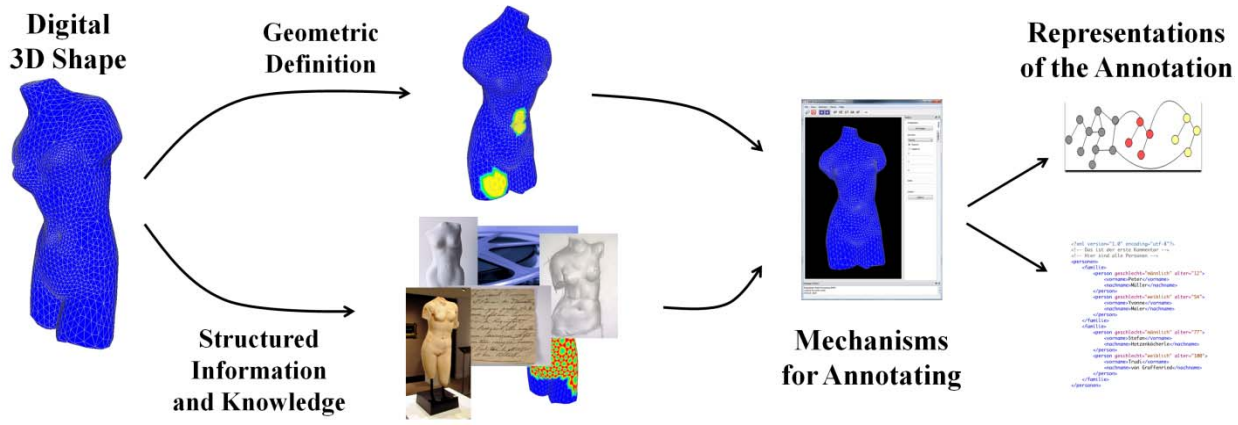

Figure 3: Building blocks of the annotation process.

A stable version of the AnnoMAD tool has also been released with all the features required by the project. The work performed during Year 3 included the extension of the tool and the implementation of new features to load, browse and display terms coming from different thesauri. The mechanism to visualize the terms also includes auto complete functionalities and the possibility to use terms coming from thesauri for query formulation and for annotations. An update method has also been included in order to deal with the possibility to manage the new terms proposed by annotations for the thesaurus.

The AnnoMAD tool has been totally rewritten in Qt/C++ to guarantee full integration with IVB browsing. The integration now provides the possibility to display URIs and relationships available in the central repository to be used for creating new annotations. The Hala Sultan Tekke dataset was used for testing the tool.

4.3 Deviation from work plan

No deviation from the work plan. The third year work plan has been fulfilled according to the DoW.

4.4 Plans for the next period

The annotating component will be consolidated during the last year of the project. The new developed interface and interaction concepts will be tested with CH professionals and included in the deployment experiments as part of the IVB (WP7).

5 T6.3 Semantic propagation

CNR will extend the MeshLab system and the annotation tool to provide features that allow the preservation/transfer of geometrically tagged annotations between different representations.

When managing semantically tagged 3D objects, one of the key capabilities is the possibility of transferring geometrically referenced semantic information between different representations of a same abstract 3D object. To allow this functionality we have extended MeshLab to support the identification between common areas in different mesh representations.

This set of functionalities allows the transfer between different mesh representations of the information that has been associated to a given portion defined as a subset of the faces of the original mesh. The implemented algorithm is based on the computation of the Hausdorff distance metric between the two representations. According to this metric we can find the regions of the mesh that are below a given distance and identify them as equivalent in the two representations. The use of the Hausdoff metric gives the user a metric non-subjective approach to the identification of equivalent parts and it allows them to keep this process under a well defined given error: e.g. portions of the meshes that are more 'different' than a given metric threshold are not selected as corresponding.

In the below illustration we show how we can transfer a semantic tagging, in this case identifying the left arm of the statue, from a lower resolution model on which was computed to a much more complex high resolution model for which a user-driven selection would have been more problematic.

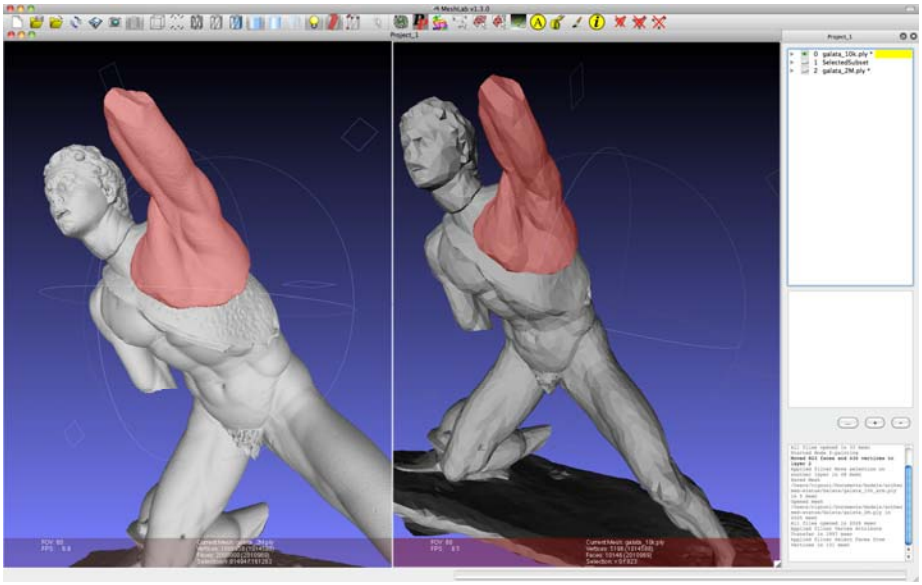

Figure 4: Semantic Propagation

6 T6.4 Co-referencing resolution tool

6.1 Work planned

The working plan for Year 3 included the release of the beta version of the Co-referencing Resolution Tool.

6.2 Work performed

During the reporting period we should have implemented the beta version of the Co-referencing Resolution Tool. However, a significant part of our resources was devoted to WP3, in order to provide a working RI to the partners since the RI is the main integration platform of most of the 3D-COFORM technical outcomes.

So, our work in the CoRef Tool has been significantly delayed. We provided the underlying mechanisms to support the desired functionality, but the implementation of the User Interface has started late and we anticipate delivery of the beta version of the CoRef Tool in Month 42 with six months delay.

6.3 Deviation from work plan

As noted above, the beta version of the Co-referencing Resolution Tool has not been delivered due to change of priorities. While we provided the underlying mechanisms to support the desired functionality, the development of the User Interface has been significantly delayed.

No significant corrective actions are required: this deviation has not had significant impact on development elsewhere in the project since the CoRef Tool is independent of other developments undertaken during this year. The lack of the CoRef mechanism might reduce the performance of the system but it does not create a malfunctioning system. As such, it does not cause dependencies with work in other work packages. However, we will prioritize delivery of this tool in the first period of year 4.

6.4 Plans for the next period

Continue with the implementation of the CoRef Tool in order to have a beta version by Month 42.

7 Publications from WP6

[T6.1.1] A. Felicetti, M. Lorenzini: "Metadata And Tools For Integration And Preservation Of Cultural Heritage 3d Information"**,** *23rd International CIPA Symposium, Prague, Czech Republic, September 12 - 16, 2011.*

[T6.2.1] Pena Serna S., Scopigno R., Doerr M., Theodoridou M., Georgis C., Ponchio F., Stork A.: 3D-centered media linking and semantic enrichment through integrated searching, browsing, viewing and annotating. In VAST11: The 12th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage (Prato, Italy, 2011).

[T6.2.2] Pena Serna S., Rodriguez-Echavarria K.: 3D Shape Annotations for Semantic Enrichment. Submitted to Eurographics STAR 2012.

[T6.2.3] A. Felicetti, F. Niccolucci: A Repository for Heterogeneous and Complex Digital Cultural Objects, *VAST2011: The 12th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage (Prato, Italy, October 18th-21st 2011)*

8 References

1. *"SKOS Simple Knowledge Organization System Primer"*, W3C, Antoine Isaac, Vrije Universiteit, Amsterdam, Ed Summers, Library Of Congress (http://www.w3.org/TR/skos-primer/)
2. *"Art & Architecture Thesaurus Online",* Getty Research Institute, (http://www.getty.edu/research/tools/vocabularies/aat/)
3. "International Standard 2788, Second Edition. 15-11-1986", International Organization for Standardization, (http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnucsnu=7776)
4. *"WebTMS System, User Manual 2011"*, ISL, ICS, FORTH.