Final Activity Report - II

for the period

July 1 2006 – December 30 2007

1. Overview of the Consortium Activities

1.1 Project Highlights

The AMASS Project main goal was to provide hardware support for a *general* search platform in databases (LCI), multimedia search based on textual annotations (VCR), and semantic search (iSOCO). We presented our *demonstrator,* poster *(Exhibitor Poster IP_7.pdf),* and a full paper *(IP07_AMASS_Paper.pdf)* describing our project at the IP 07 Meeting, which took place in Grenoble, Dec. 5-6, 2007. We consider this paper as our final Milestone.

Although we had to slightly redefine our plans and adapt it to hardware-based restrictions, the AMASS project achieved all its main goals. We find it astonishing that even after two and half years of design and FPGA programming, that the FPGA based system is (depending on the query) 25-50 times faster than the software implementation running on an Intel 3 GHz Pentium processor, although the FPGA card is running “only” at 100 MHz. These results are not caused by a poor software implementation: in fact, the C++ implementation of the AMASS processor is the same “Golden Model” as used in hardware and is using all bit-by-bit operation capabilities supported in C/C++. The large gain in hardware performance has been made possible by a very careful design and implementation of the hardware version, in which one has used the very fast (and expensive) SSRAM-ZBT memory available on the card as well as fully parallelizing the algorithm. Not a single clock tact gets lost when running the hardware!

In Figure 1 we recapitulate schematically the AMASS platform. The main goal was to allow search also with/on incomplete or noisy data. The basic software platform is defined around the Content Addressable Record Extraction (C;A;R;E;) engine, a module built and used in its applications by LCI. This provides a classical data base organization (records and fields) but does an approximate full text search based on structured or unstructured queries. This approach is used at LCI for identifying a certain client(s) from the receiver’s database based on the text extracted via OCR from paper mails or directly from electronic entry documents or emails. This quite general setting allows one to use it as the Search API in Figure 1. C;A;R;E; (the Search Engine) works in both exact and fuzzy mode. When searching noisy text, the slowest part of the search algorithm, namely selecting the most "similar" words from a data-generated dictionary has been delegated to the hardware (Common AMASS Stack).

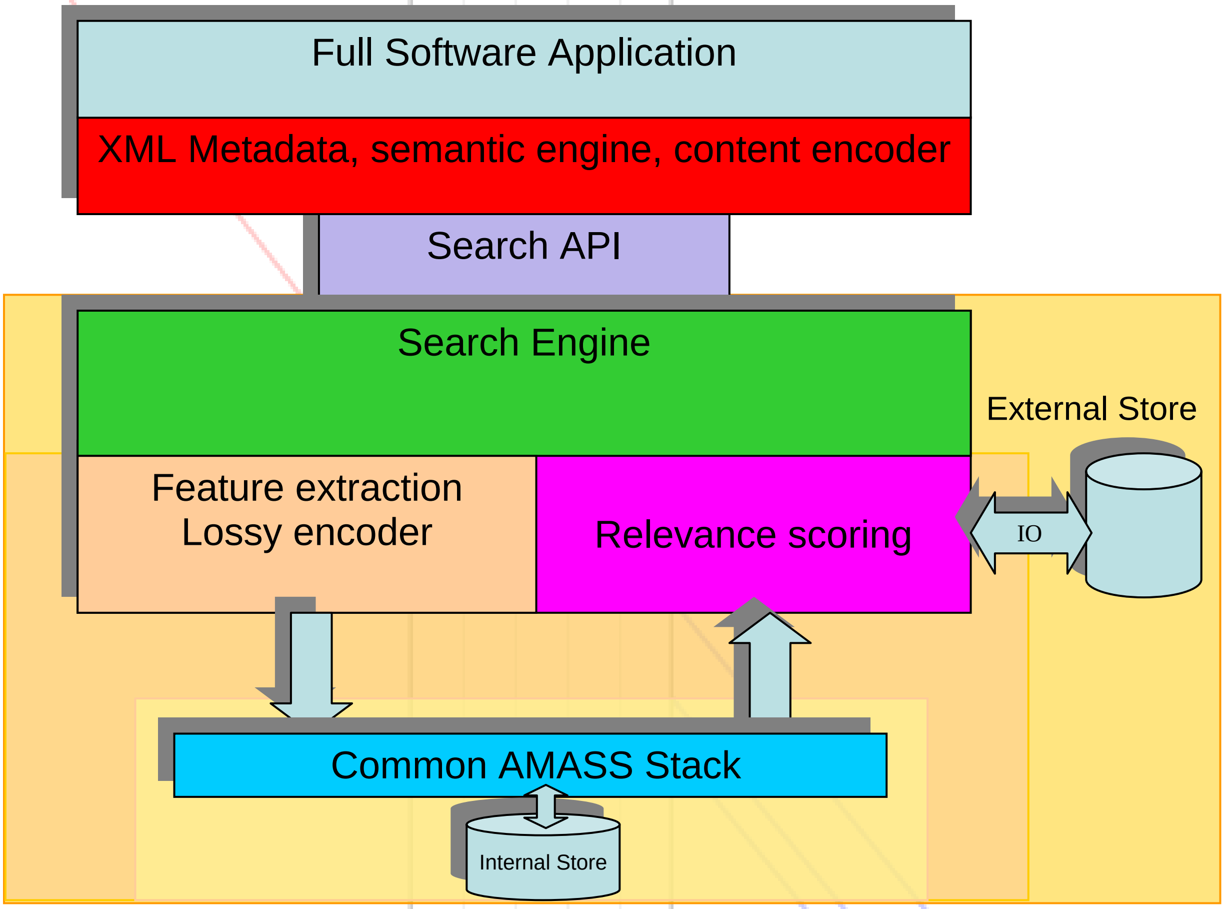

Fig. 1: Schematic representation of the search-systems as used by the SME software products. The Search Engine is provided by the LCI's C;A;R;E; engine. A simplified Search API links C;A;R;E; to the other SME platforms. Downwards, the internal store system consists of signature attribute matrices (SAMs) of the dictionary generated from the original data at indexing time. This is stored in fast SSRAM-ZBT (32 MB), searched via a new parallel algorithm ("Sequencer") and returns a sorted list of elements in decreasing order of similarity with the query. The sorting unit is also implemented on hardware.

VCR

One of the most demanding parts for the SMEs was to exchange their actual search API to the AMASS Search API. VCR uses the (exact) XML Database Tamino, a product of Software AG, in order to search for certain topics based on the text seen on video frames. VCR uses either its own developed OCR engine or manually annotated text. Since Tamino provides only exact search, VCR used to store all high probability variants of a certain OCR generated word into the Tamino database. This increased the storage requirement to the point where the XML-based search became very slow. That is not necessary with C;A;R;E; since there both the indexed data as well as the queries can be noisy. We present below the main results concerning the VCR improvements when switching to the AMASS (CARE based) platform – for details see the document "*Performance tests for applications running on the AMASS platform.pdf*".

		TAMINO		AMASS-CARE	
		Full	Annotated	Full	Annotated
Number of relevant documents found		2019	2941	2252	**3464**
Average Precision		0.12	**0.97**	0.03	0.38
Average Recall		0.52	0.55	0.6	**0.67**
Search Time (milli-seconds)	Average	50.255	49.9142	8.57	**4.994**
	Maximum	207.81	207.81	**6.25**	12.56
	Minimum	15.625	15.625	**<1**	**<1**
	Standard Deviation	97.4	97.5	**9.28**	18.484
Percent of data for which the best recall was achieved		40.88	41.89	49.26	**53.83**
Percent of data with no relevant results		31.84	31.84	25.39	**22.19**
Average time (ms) needed to identify one relevant item		29	29	7.4	**3.75**
Average time (milliseconds) spent totally inefficiently		32	32	7	**4**
Bpref		7.69987	**0.329**	10.02153	0.54415
Map		0.11205	0.54898	0.11594	**0.63657**
r-precision		0.17051	**0.9756**	0.10262	0.73996
precision@20		0.12967	**0.97332**	0.03059	0.36516
F1@20		0.08154	**0.58697**	0.03571	0.31264
AUC		0.154692	0.559599	0159075	**0.644605**
MRR		0.14173	0.6732	0.14199	**0.7439**

Best results are highlighted. The average time improvement is about 4. The small loss in precision can be attributed to the fact that the AMASS-CARE variant stores only the most probable OCR word, not all variants as in Tamino.

iSOCO

Semantic search is based on iterative exact search based on the URF scheme (http://en.wikipedia.org/wiki/Resource_Description_Framework) . The *U*nified *R*esource *I*dentification scheme consists basically of the triple (subject, predicate, object), where all three components can be URF's themselves. Some objects, however, are concrete instances represented as Unicode literal strings. Hence, a given ontology will be basically a directed graph whose leaves could be instances while other leaves are not. See "*example_NTriple_class.pdf*" for an example of defining classes and this "*example_NTriple.pdf*" for an example of defining data. Class names (as objects) are defined by the ontology and should be exact, while data objects might be incomplete or noisy and thus must be searched approximately.

Several description languages exist around this basic framework. In general, ontologies are built manually for a given domain and stored in a database. Search happens at the level of a natural language query using an appropriate parser. For instance, the question: “How old is Bill Gates?” must be parsed so as to define “Bill Gates” as an instance and “How old ?” as a time-relationship. Answers can be generated by parsing the tree with relationship type queries (exact) and instances (may be fuzzy). Today this kind of semantic search is done only exactly and this restricts its use to perfectly formed queries. Allowing the instances to be identified even with missing or noisy parts enlarges the scope and power of the semantic search. In addition, speeding up both exact and fuzzy search leads to very large gains in processing queries while searching the ontology tree. For instance, if the single query speed is increased by a factor of 2, iterating the semantic tree up to a depth of 5 would improve the overall performance by a factor of (2^q)^5, where *q* is the mean cardinality of a node and a^b means “a to power b”.

A simple example of an ontology is provided by iSOCO and can be found in “AMASS_Ontology_example.pdf”, together with the actual class and data definitions referred above. After linking the C++ AMASS API to their Java framework, iSOCO found an average search speed improvement of 24 and more on single step searches, besides extending the instance identification range of their parser. Further speed improvements could be achieved if the results would be passed via memory instead of an XML file port. LCI will provide this extension once iSOCO plans to commercially utilize CARE. The data are taken from “*Performance tests for applications running on the AMASS platform.pdf”*

		ISOCO	ISOCO-CARE
Recall	Mean	**0.7008**	**0.7008**
	Standard deviation	**0.45**	**0.45**
	Maximum	**1**	**1**
	Minimum	**0**	**0**
Precision		Undefined	Undefined
Search time	Mean	4176 ms	**163.67 ms**
	Standard deviation	22.0885	**578.71**
	Maximum	241277 ms	**5769 ms**
	Minimum	151 ms	**12 ms**
Average Time spent to retrieve one item		Undefined	154.8ms
Average time spent inefficiently (with no results found)		9220 ms (after removing the outlier it is 2390 ms)	**129 ms**

It might be interesting to read the answer system evaluation sheet attached as *ISOCO test_sentences.xml.*

LCI

At LCI we are actually developing the C;A;R;E; version 2.0, which will have, among others, the ability of automatically using the hardware support provided by the AMASS platform. LCI Research invested heavily in buying two FPGA boards (each over 7000 Eu) and hired Francois Vuillod, who did the FPGA work at the University in Ulm. We are preparing a patent application for the Golden Models and the hardware platform, which - due to internal priorities - will submitted as soon as the new CARE framework is finished. Therefore, we will ask our Referees to handle confidentially those parts of our Reports noted as such, in order to not invalidate the patent application.

In parallel with embedding the AMASS results into our product, we also re-evaluated some of our algorithms and design in order to support seamlessly the hardware. Among other things, the new CARE engine will support 64 bit platforms, multithreading, and hence modern multi-core processors. Full UNICODE 16 and 32 bits will be also supported and thus a new range of application possibilities. We also want to explore the possibility of using GPUs instead the FPGA platform. From a commercial point of view, however, producing our own FPGA cards is the preferred approach. The FPGA part will be designed in Kirchzarten and the boards and packaging will be done at Kofax Inc., Irvine, CA, USA.

The new CARE engine is expected to perform 4-10 times faster on fuzzy search and is intended to handle much larger databases then previously. It will be sold as a server-like web-service environment or as a module used by the Kofax Transformation Modules. The decision whether to develop a custom FPGA card or not, will be taken around July 1st 2008.

Web-presence

Our web presence can be found at www.amass-platform.com. When trying to put inline some of our demonstrations we encountered copyright and licensing problems for the most interesting, the VCR video-frame search based on text queries. We do have, however, running demonstrations of database search, video-search, direct FPGA based vs software based filtering, semantic search, and protein search at LCI's headquarters in Kirchzarten, Germany and would welcome the visits of our Referees. UAB offers the possibility of accessing the video application and the online AMASS GUI for testing the FPGA card by remotely accessing the Windows server 158.109.70.122. User: amass_user Passwd: cephis2007. The AMASS GUI demo is at : http://158.109.70.122 and the video demo at http://158.109.70.122/VideoSearch.

1.2 Participants and Consortium

The participants in the AMASS Consortium were

SME

1) LCI GmbH (LCI) Kirchzarten, Germany, Co-ordinator – automatization of document processing
2) Intelligent Software Components (iSOCO) Madrid, Spain – semantic search
3) Visual Century Research (VCR), Barcelona, Spain – multimedia management systems

RTD

4) University of Ulm (UniUlm), Germany – hardware development
5) Universitad Autonoma of Barcelona (UAB), Spain - embedded software and hardware
6) Kepler-Rominfo (Kepler), Bucuresti, Romania – Standard verification procedures
7) University of Iasi (UAIC), Romania – software technology

The AMASS Consortium Agreement has been worked out and signed in July 2005 by all participants. The EUC has received one of the eight original exemplars.

1.3 Status Changes

In February 2006, DICOM Group Ltd. (www.dicomgroup.com) has acquired 100% of the owner rights of LCI GmbH, where it has previously owned a 19% share. The LCI GmbH remains a separate entity, without important changes in either its leadership or staff. The Consortium partners and the EUC Officers have been informed about this change. LCI GmbH has applied at the EUC for remaining the Co-ordinator of the Project, despite the fact that the DICOM Group is not an SME according to the EU definition. In February 2008 the name DICOM Group Ltd has been changed to Kofax Ltd and LCI GmbH has been renamed Kofax Development GmbH. The Author decided to refer throughout this Report to LCI instead of Kofax Development GmbH, because the name change is not yet confirmed by the Handelsregister Freiburg.

In August 2007 Prof. H.J. Pfleiderer, leading the Microelectronics Chair at Ulm University retired. The members of his group are now working at different prestigious research institutions (IBM Deutschland, First Fraunhofer Sankt Augustin, etc.). Dr. Juan Maria Sanchez, the leader of the project at VCR has recently left the company.

1.4 Short summary

This Final Report includes several deliverables, milestones and ad-hoc documents making the follow-up of the AMASS development easier to follow. The main scientific and technological goals achieved during the AMASS Project are the following:

- Creation of the “Golden Models”, C++ programs describing the new algorithms, which take into account the additional possibilities offered by the hardware (delivered on Periodic Report 1).
- Definition of the “high-level” API, connecting the AMASS top software module to the three SME application architecture. This API defines the common calls supported by the search engine at the host level.
- Definition of the “low-level” or “hardware-level” API, through which the host and the FPGA communicate. This is actually a C/C++ API, which is supported on the hardware side by a driver (controller) and split into hardware-dependent micro-operations. On the software side, the hardware is simulated

by a general purpose C++ Golden Model implementation, simulating as close as possible the hardware functionalities.

- Converting the C++ libraries into SystemC for a better understanding and simulation of the hardware architecture. May be the most critical issue addressed here is how to solve the “bottleneck” problem of accessing the memory as fast as possible. After several iterations with the group in ULM we decided to use only the very fast SSRAM-ZBT (32 MB) available on board. That will restrict the use of the hardware support to a “word-dictionary” based approach. For normal uses, the available 2 million word entries should be enough to cover most applications and data.
- Definition and implementation of different performance measures for text information retrieval, semantic search retrieval, and multimedia annotation retrieval, in order to be able to measure both the quality of the search as well as the technological parameters and improvements due to the hardware support. While for text retrieval such measures have been already standardized by “experimental” conferences like TREC, this required new insights into how and what to measure for both semantic web type and multimedia management applications.
- For standard verification and performance measures a set of “Golden Data” has been defined, on which the tests must be performed. This has been provided by the three SMEs in their own application fields.
- A realization of the embedded software search engine for both the Golden Model and different applications of the three SME has been demonstrated at the *IP 07 Meeting* in Grenoble. They are also available at LCI and UAB. We will try to make some of these directly accessible via Internet but this has some limitations because of copyright and licensing issues. This is also true for the most spectacular application, the VCR demonstration of searching video-streams of news with textual fuzzy matching.
- Two new possible application fields have been proposed by VCR (content based image search) and LCI (genomic search and biometrics applications). Among these, the genomic search is in the most advanced state and could be moved to the (scientific) market in short time.
- A patent application by LCI will be soon submitted to the European Patent Office. This includes the full LCI StringOverlap procedure, the LCI Sequencer algorithm and its FPGA based hardware implementation. As applications we will cover the general platform framework and the specific database search.
- The LCI is now building the results obtained during the AMASS project into its next product, the Kofax Transformation Module version 4.0
- The work on the AMASS project provided students and young researchers the opportunity to address some of those problems in their Diplomarbeit and PhD Thesis. We include for information also these documents (only in electronic form on the CIRCA server).

On the management side, the most important tasks have been:

- Performing the negotiations with the EUC on the actual contract,
- Negotiating and signing the Consortium Agreement,
- Distributing the EUC contribution between the participants,
- Organizing the different regular Meetings of the Consortium,
- Co-ordinating the technological and scientific work
- Being a bridge between the Consortium participants and the EUC Officers in different administrative issues. This was necessary especially since both the

Co-ordinator and several Consortium members are involved for the first time in EU projects. Thanks to our EU Officers for their patience!

- Writing the Periodical reports
- Designing and implementing the project's web page at (www.amass-project.com)

1.5 Meetings and communications

The Consortium has often discussed the actual tasks, problems, and solutions at different Meetings.

The general AMASS Meetings have been held at:

- Ulm, Germany, 4-5/07/2005
- Barcelona, Spain, 9-11/08/2005
- Kirchzarten, Germany, 22-23/11/2005
- Madrid, Spain, 27-28/02/2006
- Bucharest, Romania, 05-06/07/2006
- Ulm-Kirchzarten 10-11.05.2007
- Barcelona: final Workshop Oct. 1-4, 2007

As the project progressed, it turned out that it is easier for all participants to use directly public domain software for exchanging and storing source code, for example. Therefore, we have not implemented on the AMASS project web-server (http://www.amass-platform.com) the intranet part and thus spared a lot of time and costs for its implementation. The embedded server runs can be accessed via Remote Windows access at //158.109.70.122. See userids and passwords above.

2. Post Mortem Analysis

In good tradition, once a project is finished one has to consider the difficulties encountered with the goal of learning how to avoid them in the future. We call this the "post-mortem" evaluation of the project.

- On the hardware front, the original plan of attacking the problem along two different strategic lines, a traditional one based in ULM and a more abstract one at UAB, did not work out properly. While the actual implementation comes from ULM and the low API conversion from UAB, a lack of stronger interaction and differences of perspective impeded the development of our own PC card at UAB. Although we had a very good start, the lack of leadership in deciding hardware development by the Co-ordinator, who took real differences for semantic ones, left for too long important issues being heavily discussed but not being addressed.
- On the SME part, LCI has been too slow on providing the CARE libraries with the high level AMASS API wrapper, so that the other SME's had perhaps not enough time to build it timely into their respective platforms. Technical and

internal organisational changes at both LCI and VCR did not help either with the priorization of research work. These are but general problems occurring at very small companies struggling either to survive or expanding too fast.

- On the positive side, we made very good experiences with both Romanian partners, both in work quality, delivery timing, and organisational skills as proved by their excellent contributions to this Report.
- Finally, finishing the work and these Reports has taken much longer than planned or expected by the Co-ordinator. The Project Co-ordinator takes all the responsibilities for the negative issues discussed here. While he is extremely grateful to his EU officers for their support and understanding during the whole Project, the fact that in the last two and half years we had three different scientific and two different financial officers did not contribute to a stable communication base with the EC.

3. Deliverables and Milestones

3.1 Deliverables and Milestones

A list of delivered documents is attached in Appendix A below. An electronic copy has been deposited on the CIRCA Server. The AMASS public web-page can be found at http://www.amass-platform.com, where a slightly extended version of this Report will be published.

The Periodic Management Report II

and

The Plan for Using and Disseminating the Knowledge

are attached in separate documents.

Appendix A: Attached documents

Deliverables in Report form:

Deliverable 1.2: Extending the AMASS Platform.pdf

Deliverables ULM (WP 3.x) Confidential

- UAIC report.2006.july Golden Model.pdf
- Sequencer FPGA implementation.pdf
- MemIF_refman_ULM.pdf
- hw_api3 UAIC.pdf
- HW Qt Interface ULM.pdf

- DMA_analysis.pdf
- FV enssat_report.pdf (Diplomarbeit – only CIRCA Server)
- DA-JS.pdf (Diplomarbeit – only CIRCA Server)

UAB:

AMASS_Deliverable_4_1 .pdf
AMASS_Deliverable_4_2.pdf

KEPLER-ROMINFO:

AMASS Deliverable 5 1.pdf
D5.1- Standard verification procedures v03 20060323.pdf
AMASS Deliverable 5 2.pdf
D5 2 - Performance Measures Test Methods v01 20060310.pdf
AMASS Deliverable 5 3.pdf
AMASS Deliverable 5 4.pdf
D5 4 -Performance tests for applications running on the AMASS platform - v04 20080212doc.pdf (Final Milestone)

Milestone IV: Public

Exhibitor Poster IP_7.pdf
IP07_AMASS_Paper.pdf

Semantic Search ISOCO examples (Confidential)
AMASS_Ontology_example.pdf
example_NTriple_class.pdf
example_NTriple.pdf
ISOCO test_sentences.xml

Periodic Activity Report II.pdf (this report)

AMASS Project Deliverable status

Code: red: late but not deleted | green: delivered | yellow: being submitted

Deliv	Title	Delivery month	Nature	Dissemination level
D1.1	Report and Software: "Common API and Golden Models for raw, multimedia and internet data"	6	O	RE: Milestone I
D1.2	Report and Software: "Golden Models for feature encoders"	20	O	RE
D2.1	Report "SoC Architecture and components for the AMASS Platform"	12	R	RE: Milestone II
D2.2	Embedded commercial platform for raw text search in databases	20	O	CO: Milestone II
D3.1	Iterative Hardware platform design methodology	16	O	RE
D3.2	The AMASS Platform	26	P	PU
D3.3	Virtual Component Representations and Codes	24	O	CO
D4.1	Interface Selection for External I/O	9	R	RE
D4.2	Fast FPGA-embedded hierarchical memory system	12	O	CO
D4.3	Report "System Architecture and Applications Design for the AMASS Platforms" - removed	18	R	RE: Milestone III
D4.4	Patent submissions	30	R	RE
D5.1	Technical Report: "AMASS standard verification procedures"	4	R	PU
D5.2	Report "Performance Measures, Test Methods"	6	R	RE
D5.3	Report "Performance testing tools and visualization"	16	R	RE
D5.4	Report: "Performance tests for applications running on the AMASS platform"	26	R	RE: Milestone IV
D6.1	Consortium Agreement and constitution of Consortium Board	3	O	RE
D6.2	Project management meeting reports, including the kick-off	0,5,17,23	R	RE
D6.3	EC reporting period required reports and cost statements, including the Mid-term assessment report.	12,32	R	RE
D6.4	Project web site and applications	6	O	PU
D6.5	Knowledge dissemination reports I and II	12,32	R	PU
D6.6	AMASS Demonstrator and Workshop	30	D	PU: Milestone IV

AMASS