



Project Number: 032518

QVIZ

Query and context based visualization of time-spatial cultural dynamics

Specific Targeted Research Project

Information Society Technologies

State of the Art D2.1

Due date of deliverable: 30/09/2006

Actual submission date: 30/09/2006

Start date of project: 01/05/2006

Duration: 24 month

Umeå University

Revision 2

D2.1 State of the Art Report

Number and name	D2.1 State of the Art			
Workpackage	WP2			
Task	T2.1			
Date of delivery	Contractual	30/09/2006	Actual	30/09/2006
Code name			Version	draft <input type="checkbox"/> final <input checked="" type="checkbox"/>
Nature	RE+ online wiki			
Distribution Type	Restricted (<i>to Consortium</i>)			
Authors (Partner)	<p>Bob Mulrenin, bob.mulrenin@salzburgresearch.at Fredrik Palm, fredrik.palm@humlab.umu.se Kalev Koppel, kalev.koppel@regio.ee Indrek Kuben, Indrek.kuben@ra.ee María Isabel Díez del Val, midiez@tid.es Patrik Svensson, patrik.svensson@humlab.umu.se</p> <p>Contributions to the SOTA Electronic version by the above partners and the following: Andrén Peder, sveu11@svar.ra.se Humphrey Southall, Humphrey.Southall@port.ac.uk David de Fco, qviz@tid.es Nikoletta Czako, nikoletta.czako@humlab.umu.se</p>			
Contact Person	Patrik Svensson, patrik.svensson@humlab.umu.se Fredrik Palm, fredrik.palm@humlab.umu.se			
Keywords List	State of the Art Report, technologies for archival organisation, social software, gazetteers technologies, GIS technologies, visualization, community and knowledge repositories.			

Table of Contents

QVIZ	1
STATE OF THE ART D2.1	1
A. EXECUTIVE SUMMARY	4
B. SOTA ELECTRONIC VERSION	4
C. METHODOLOGY.....	4
E. SOTA STRUCTURE AND ANALYSIS	7
D. THEMES, HIGHLIGHTS AND ANALYSIS.....	7
1 <i>Introduction</i>	7
2 <i>Archive and content organisation</i>	7
3 <i>Technologies relevant to QVIZ</i>	9
4 <i>Ontology concepts, knowledge contextualization and related standards</i>	10

A. Executive Summary

As a project involving social software, we chose to work with social software tools to communicate, record and organise much of our work. A State-of-the-Art report is best utilised as a resource if it is a "living document" and as we progress in the project, new materials can be integrated when we wish to communicate them to other partners. Consequently, the State-of-the-Art report is an electronic online document and will be updated periodically to communicate new or updated materials relevant to our work. This document gives a summary of the results and an overall contextualization and analysis. A snapshot version of the electronic version has been provided as an appendix.

B. SOTA Electronic Version

The SOTA Electronic Version is available online at:

<http://qviz.humlab.umu.se/index.php/SOATopicIndex>

Accounts provided upon request to Fredrik Palm fredrik.palm@humlab.umu.se. For convenience we provide an appendix with a pdf-based snapshot of the SOTA at the time of this deliverable. It is important to point out that this version is a flat, textual version of the dynamic and hypertextual electronic version.

C. Methodology

In order to work collaboratively, partners have interacted using various internet based tools; mostly through email, Skype, and the Wiki.

In general, an important goal of SOTA is simply to support our tasks in the workpackages and even to feedback into the SOTA as we progress deeper into various tasks of other workpackages. For example, as we understand better the interests of the QVIZ Communities of Practice in the user trials and document these in WP3 task 3.3 Domain Ontology, we can update the SOTA with additional relevant materials. But still it was necessary to carefully review, select and filter materials before writing to the wiki online presentation. Partners have currently many offline notes and work that will remain offline until needed because too much material can distract partners when that is not needed.

Initially the SOTA topics were created as a table of contents in the wiki, and contributors could then augment and rework the index over time, add new topics or consolidating them. As our work progressed during requirements capture and discussions of WP3 issues (to stimulate requirements capture and communicate our ideas to one another), we tailored

some of the contributions to the SOTA. The following table reflects some of the driving issues that led to our topic definitions and contributions.

	Driving issues
WP2 Requirements	<ul style="list-style-type: none"> • Enhance requirements and ontology requirements capture by explaining some of the problems the project and current SOTA, especially relating to administrative units and administrative units and GIS issues. • Help to describe the archive domain, organisation and business processes to help understand requirements relating to use and access of archive materials. Support initial discussions of tentative access methods relating to requirements and WP3. • To gain knowledge about existing GIS-technologies, that might help capture both user and technical requirements, for example GIS-clients, servers, tools for visualisation. • Need to discuss existing projects and system for handling administrative units ontology, for example using the Vision of Britain as a base-line example. • Based on existing project and system, there was a need to better understand how different approaches for knowledge visualisation can make new user and technical requirements.
WP3 Tasks 3.1 Selection of Content	<ul style="list-style-type: none"> • Partner archive descriptions, organisation and basic archive types that help us define target archival materials. • Explain to developers how archives are organised in general and more specifically, the possible structure of inner organisation of archival materials in different institutions. • Digitisation practices: selection of resources, use of metadata and how are these related to the general archival description. • Specify initial QVIZ-related archival content and analyse search opportunities of different databases, including given hierarchy of administrative units. This provides enough requirements which will followed up in more detail in WP3.
WP3 Tasks 3.2 Administrative Ontology	<ul style="list-style-type: none"> • Partners need to understand the complexity of administrative units in a European context, since a cross-border approach results in additional complexity. • Partners also need to be aware of previous initiatives for data modelling administrative units, for example the notable problematic and expensive GIS-centric solution, which some past initiatives have invested largely into without obvious success. • It is obvious that that archives have very large amounts of

	geographical knowledge held within existing systems and one important point concerns making use of these data without requiring either major changes to software or the addition of coordinate data.
WP3 Tasks 3.3 Domain Ontology	<ul style="list-style-type: none"> • The need to analyse archive thesauri formats relevant to enhancing archive materials with semantic descriptions • Sub-domain: common knowledge organisation structures? • Community related research <ul style="list-style-type: none"> ○ Discourse ○ Social software related ○ Relevant archival thesauri and standards ○ Archival and record keeping standards • Document /metadata standards relevant to digitisation • The importance of evaluating and interacting with relevant projects in this area. • Basis for QVIZ community ontology and basis/format for thesauri.
WP 4 Specifications	<ul style="list-style-type: none"> • What core components might we consider to adapt/reuse in our system architecture? • Community related research and technologies and their influence on overall architecture • Need for knowledge about current GIS technologies, both on web client side and server backend. • Better understanding of methodologies for GIS-based modelling of dynamic visualisation
WP 5 & 6 - implementation class	<ul style="list-style-type: none"> • Technologies useful for implementation, such as social semantic software, ontologies repositories, client and web components, protocols for information exchange of archival and user-based information, data models/repositories for administrative unit data etc.
	<ul style="list-style-type: none"> • State-of-the-Art reports, relevant projects

Table 1 Driving Issues

E. SOTA structure and analysis

The top level structure of the SOTA electronic document includes the following categories:

- General resources
- Archive and content organization
- Technologies relevant to QVIZ
- Knowledge related (Ontology, thesauri, etc) or standards

Each of these main categories subsumes a large set of documents and resources.

D. Themes, highlights and analysis

1 Introduction

This section provides a thematic account of some of the major areas covered in the electronic SOTA. For each theme we list a number of key points that are relevant to future work in the project as well as to an understanding of the state of the art in these areas.

Important general sources for this work include information about a range of related projects and initiatives, relevant email lists, information about current development and policies, other SOTA reports, material on software tools and systems, and project-internal feedback loops.

2 Archive and content organisation

Within this theme, the structure of administrative units at partner sites was initially specified at a high level. Analyses of organisational schemes for archival documents provided an important foundation as well as an overview of existing standards for describing archival resources and discussion of requirements for semantically enhanced resources.

Another important focus within this theme was the description of different approaches to visualizing and modelling administrative unit data. This work included researching various methodologies for providing highly functional interfaces for users with limited knowledge about archival structures.

Here are some examples of important topics covered in the electronic SOTA:

- Understanding of approaches to handle the complexity of Administrative Units, such as ADL (Alexandria Digital Library) Gazetteers and ADL feature type thesaurus.

- Communication of models and structures of existing systems, primarily the Vision of Britain Gazetteer approach for administrative units. That information was necessary to understand what further development was needed for a European wide approach and also meeting the requirements for dynamic queries and visualization of archival and community knowledge related to administrative units.
- Existing GIS-approaches help to capture important user and functional requirements, such an existing system at national archive of Estonia, and also pioneering examples from leading GIS developers.
- Partners also communicated other prototypes with non-conventional approaches for access to time-spatial administrative units, using "time-less access to time-spatial resources", which help capturing user requirements and also functional requirement for a European administrative ontology.
- Description of data structures of existing data and alternative approaches for more query-flexible approaches, not just simplicity and performance.
- Relevant to the work on archive organisation and the content selection phase, we explored archives and archive organisational perspectives. Fond-level organisation at the relevant archives was explored. This is relevant to the QVIZ domain ontology modelling and how we work with each archive in WP3 T3.1.
- Access methods were contextualized and discussed. The tentative access method addressed turned out useful for requirements discussions about visualisation, institutional user interface integration, and archive materials access and search results.
- Archival standards such as ISAD(G) (General International Standard Archival Description) are useful for organizing queries, visualization and QVIZ reference material and metadata for social semantic structures. Furthermore EAC needs more investigation, although we have seen few successful use of that standard.
- The SOTA has helped to investigate access strategies involving social book marking, semantic descriptions, archive descriptions and archive material registration data.
- METS is interesting for digital library formats, and Recordkeeping concepts might provide QVIZ with a good background for the domain ontology and archive repository model. One possible connection would be in context with METS and the Fedora digital repository.
- To understand the challenges of archive digitized objects, we found certain image formats in common use, such as DjVu. This format will need additional supporting components to handle semantic annotation, especially to augment an existing DjVu web browser plug-in. JPEG formats should be also supported in QVIZ because the web technologies and techniques are adequate for enabling annotation of images or image segments via a web browser without special plug-in.

3 Technologies relevant to QVIZ

This theme aims at initially describing possible technical baselines that can be useful within different aspects of the project. There are many different digital tools for social collaborative processes and for QVIZ it is necessary to be aware of these and evaluate their potential. Furthermore, QVIZ will need technologies for handling different types of archival and community related repositories. Other important technologies involve tools for handling dynamic and rich user interfaces, including map and complex query mechanisms, using ontologies of European administrative unit, archival organisation as well as community added semantic description.

3.1 Social software

- Supportive technologies for FOAF such as FOAF maps, folksonomies, folksonomy based representation (such as tag clouds) might lead to interesting search capabilities and powerful visualisation of communities, persons, projects. If we look also at relevance and similarity measures, users might discover and visualise relevant persons or projects particular to their interests (via semantic query/filtering). Such an FOAF enhanced approach should also aim at integrate into the domain ontology design, especially to study person related classes. User profiles can possibly be based on FOAF such as FOAF-REALM and follow-up projects.
- Knowledge about different kinds of aggregation, subscription and information technologies are important to QVIZ. This includes news feed style (RSS aggregation) subscription or advertisement of QVIZ resources and CoP activities. A news channel could be simply a "query" into QVIZ for certain resources, subjects, topics, saved map /knowledge queries, etc.
- Agent frameworks have been investigated and one platform looked at is the Jadex framework. One key question addressed here whether agents can be used to better support communities.
- Different kinds of social software have been evaluated including blogs, wikis social bookmarking and web 2.0 applications (such as Google Earths time series function).
- A number of ideas and concepts have been documented in relation to semantic wikis. Primarily, there are fundamentally different approaches in the early semantic wikis that QVIZ can learn from and decide how to make a more suitable approach for our needs.

3.2 Archival content and social semantics repositories

- Fedora Digital Repository might be relevant depending on the course of system specifications and requirements; and because of its prominence in digital repository projects. An additional repository server might also include Apache Jackrabbit because of its modelling flexibility and supporting infrastructure.
- The SOTA has helped to investigate access strategies involving social bookmarking, semantic descriptions, archive descriptions and archive material registration data

- Open source archive solutions for managing, searching and presenting archives have been investigated. These might provide an additional test implementation or event component for supporting finding aids. XTF is one such tool that has widespread use now in the US.
- The Open Archives Initiative (Dublin Core) provides an interesting OAI interface for harvesting as a means to either support external harvesters.
- Persistent object identifier technologies have been analyzed including the Archival Resource Key. Persistent object identifiers might be addressed for any registered QVIZ resource
- Our study of workflow and service oriented architectures (BPEL and SOA) showed that there are many new resources, but tools lack maturity and acceptance.

3.3 Query and visualisation tools

- Facetted "semantic" search can and should be adopted (subject, title, topic, folksonomy cloud, admin units, other relation facets)
- SOTA helped to give an overview of WebGIS functionalities, visualization issues and problems as well as a technological framework and service standards of modern web mapping. Examples of different applications were essential for requirements capture and showed what have been done in other projects. As the general tendency WebGIS is adapting the functionalities of Desktop GIS programs in their flexibility and richness including analytical tools and advanced visualization.
- Web 2.0 technologies were further investigated including advanced client and visualisation support using AJAX and FLASH based GIS tools
- Tasks and requirements discussions regarding issues such as image annotation and data representation have lead to an understanding of what web technologies might support QVIZ, in creating a dynamic collaborative environment for query and context based visualization of rich archival resources and community added content.

4 Ontology concepts, knowledge contextualization and related standards

This theme describes useful approaches and needed enhancements for the QVIZ implementation. This involves both building a good research basis for understanding the complexities of the QVIZ ontology(s) as well as tools and approaches to semantically describe archival, or other cultural heritage resources references, social objects communities and their resources and activities, content objects etc.

- There are several in the SOTA electronic version describing complexity, such as naming of different regional districts in different countries.
- We have also investigated and analysed possible methodologies to handle a European Administrative ontology, such as ADL Gazetteers, and Vision of Britain extension of the ADL-feature type Thesaurus. There are clear benefits with a Gazetteers approach but more research is need to handle the European dimension issues of multilingualism, and of course the complexity over time.

- Issues addressed include performance issues of storing large administrative units and description of different types of administrative units and possible relations in order to meet requirements discussed in the initial user requirements,
- Descriptions about how to related geographical data to Administrative ontology using for example OGC-technologies and the GML-protocol.
- One obvious focus has been the semantic wiki "Ikewiki" - Any knowledge resource (a wiki page) can have multiple classes assigned to it and the properties of the classes can be used to annotate the wiki text. What is important to QVIZ is also that one resource can be assigned to multiple classes (users assign multiple "types") thereby enhancing a search engine e.g. a SKOs concept "manuscript" might also be assigned to a BibText ontology class "Publication". Furthermore, Folksonomy tags can also be "typed" as desired by the users. Therefore it is important to better understand the relationship of typing concepts from thesauri and folksonomies - especially to understand the benefits of searching, browsing and visualizing the interrelated resources.
- A goal was to identify a common formatting standard for thesauri; to identify at least one extensive existing archival thesaurus that utilises SKOS format in some way, such as an exchange format or even as basis for a system, and to identify existing schemas relevant to potential CoPs using this thesauri format.
 - Rich and diverse SKOS based thesauri exist for learning, archives, etc. UKAT archivist thesauri demonstrate how QVIZ knowledge works might create CoP based or CoP shared thesauri based on SKOS, to create micro thesauri (sub-schemas) to create alternate structures, etc, but based on one format. Furthermore, there is evidence that we can also further "type" thesaural concepts according to other supported ontologies and this way enrich the means of discovery in semantic search facets.
 - Also, continuing with SKOs, we identified a number of possible SKOS thesauri that might be useful for any sub-domain desired by CoPs.
- One type of knowledge visualization investigated was the "Fresnel Lens" technology was investigated. This technology is interesting and we will continue to monitor its development (e.g. through the SIMILE Fresnel mailing list). As an example, Fresnel Lens might be used for visualising parts of the Community ontology, a social object or any discourse over a social object (image segment etc), FOAF object.
- Community and social objects - Discovered the potential value based on SIOC, and further conceptualisation from COIN, SWRC, AKT Reference Ontology (Support, Reference modules) and DILIGENT (discourse ontology).
- A central concept described in the electronic SOTA is social objects. Anything can be typed as a social object, we will work with semantically enriched objects, therefore we might call them "social semantic objects". The reference to any archive material social bookmark (a reference with archival metadata), an item that is the part of social discourse, an image map, an object with wiki or blog properties, or both, etc. We must be careful not to be tied to the current concepts found in the implementations of wikis and blogs, and consider only that we have social objects with particular properties that define them and where users might define how they are used, organised and visualised.
- There are articles in the electronic SOTA describing organisation of Collaborative teams (CoPs) and combinations of professional and non-professional users. A CoP

is broadly defined and there are several categories such as Community of Interest. We should ask ourselves which categories are more relevant to a particular "CoP" and include them in a semantic description of the CoP.

- Overview of ontology schemas for supporting Collaborative knowledge processes. There is possibly a need for finer granularity for specialized subgroups. CoPs can share many core ontology schemas and we need to interrelate them. For thesauri, there must be a mechanism to interrelate concepts from different schemas and provide query support. At a higher level, CoPs must be aware of what other CoPs have done and then try to use the work and collaborate together with other CoPs. Another relevant aspect covered is addressing sub-domains for particular CoPs, such as those involving archivists, or eLearning, or content objects requiring supporting ontologies.
- Initial knowledge repository proposals should consider and weigh critical issues (1) scalability and (2) performance (3) reasoning support. There are possibilities regarding RDF based relational, object relational repository solutions. Also of interest are object relational solutions such as OpenLink, RDF based such Sesame2 or OWLIM (LGPL) on top of Sesame 2 (not officially released) but has the drawbacks of using a main memory model. BIGOWLIM will require a commercial license, but is scalable without posing high demands on main memory (RAM). Since these technologies do restrict the ontology language (such as Owl-Lite subset, RDFS), this would impact WP3 Task 3.3 Domain ontology. There is little hope in the next few years to find a reasonably scaleable and well-performing knowledge repository that supports OWL-DL, for example. For that reason, we will continue our technology watch for semantic repositories that support the level of inferencing needed, that support SPARQL as a common query interface, and are reasonably scalable. Our work should also consider distributed queries over SPARQL interfaces and even model smaller knowledge nodes rather than growing a large central knowledge repository - this must be studied and discussed more.
- A Collection description (ontology) is needed because the concept "collection" can be considered fundamental to internally organized archives material references from archive descriptions and any virtual organisation, and other resources resulting from social or content management activities. RSLP is already based on RDF and has a strong following, including Michael, but we look to Dublin Core collection work and QVIZ needs for additional conceptualisations. Creating a Michael collection description from a QVIZ conceptualisation is preferable to actually basing our collection on Michael. QVIZ will need to address how and what it exposes to external services such as Michael, OAI, web based news or RSS feeds. Furthermore, the ontology supporting community/social activities (e.g. SIOC, etc.) must be further studied in context with the collection description design.